# Focal Visual-Text Attention for Memex Question Answering

Junwei Liang[1] Lu Jiang[2] Liangliang Cao[3]
Yannis Kalantidis[4] Li-Jia Li[2] Alexander Hauptmann[1]
[1]Carnegie Mellon University [2]Google AI [3]University of Massachusetts [4]Facebook Research
{junweil,alex}@cs.cmu.edu, {lujiang,lijiali}@google.com, llcao@cs.umass.edu, yannisk@fb.com

**Abstract**—Recent insights on language and vision with neural networks have been successfully applied to simple single-image visual question answering. However, to tackle real-life question answering problems on multimedia collections such as personal photo albums, we have to look at whole collections with sequences of photos. This paper proposes a new multimodal MemexQA task: given a sequence of photos from a user, the goal is to automatically answer questions that help users recover their memory about an event captured in these photos. In addition to a text answer, a few grounding photos are also given to justify the answer. The grounding photos are necessary as they help users quickly verifying the answer. Towards solving the task, we 1) present the MemexQA dataset, the first publicly available multimodal question answering dataset consisting of real personal photo albums; 2) propose an end-to-end trainable network that makes use of a hierarchical process to dynamically determine what media and what time to focus on in the sequential data to answer the question. Experimental results on the MemexQA dataset demonstrate that our model outperforms strong baselines and yields the most relevant grounding photos on this challenging task.

**Index Terms**—Photo albums, question answering, vision and language, focal attention, memex

## 1 Introduction

A typical smartphone user may take hundreds of photos when on vacation or attending an event. Within only a few years, one ends up accumulating dozens of thousands of photos and many hours of video, that capture cherishable moments from one's past, such as weddings, family gatherings or birthday parties. As discovered in [21], people may use personal photos/videos as a mean of recovering pieces from their memories about these events.

A natural way to recall the past memory is by asking questions, *e.g.* when did we last visit the zoo? To tackle this challenging problem, we propose a new VQA task named *MemexQA*[1] [2] [3], a new task for question answering on personal ***photo albums***: given a sequence of photos from a user, the goal is to automatically answer questions about the past events captured in these photos. For example, in Fig. 1, the input to a MemexQA system is a question and a sequence of photos (images + metadata) ordered by creation time, and the output is of a text answer and a few grounding photos that justify the answer. The grounding photos are necessary as not only are they useful in quickly verifying the answer but also they provide vivid information to refresh user's memory about the asked event.

Over the past few years, a number of datasets have emerged to promote research on the Visual QA (VQA) [3], [4], [14], [16], [20], [24], [25], [45], [50], [56], [57], [58], [60], [64], [65]. However, none of them is suitable for our MemexQA research due to the following key differences in the problem setting. First, our input is a sequence of personal photos as opposed to a single image or video. Although some datasets [20], [50], [56] may contain a sequence of images, these images are typically from a single video or topic. Answering MemexQA questions, however, requires reasoning over more than one photos spanning across multiple topics or events. Second, an important piece of our output is the grounding photo that can justify the answer. However, existing datasets mainly focus on the answer accuracy whereas usually do not provide such grounding photos. Third, MemexQA is a multimodal QA task. Naturally, a personal photo is associated with rich metadata such as the time, title or GPS information. Some

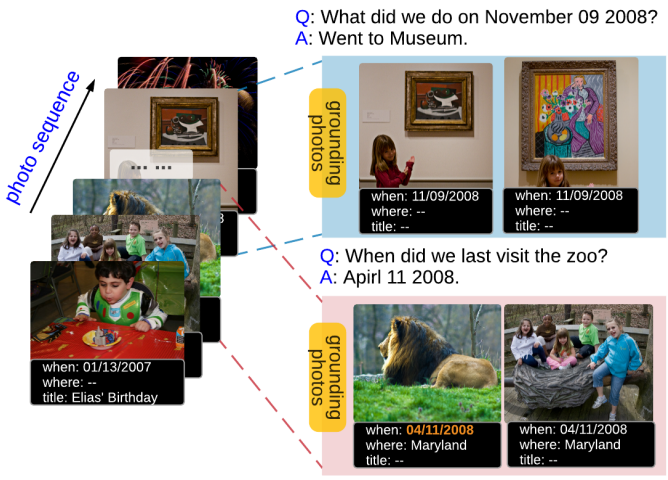



Fig. 1: MemexQA examples. The inputs are a question and a sequence of a user's photos with corresponding metadata. The outputs include a short text answer and a few grounding photos to justify the answer.

- *Lu Jiang is the corresponding author.*

1. The term *Memex* was posited by Bush in 1945 [6] as an enlarged intimate supplement to an individual's memory.
2. Dataset and models are released at https://memexqa.cs.cmu.edu
3. This work improves and concludes the studies in [22], [32]

questions need to be answered together by the photo content and metadata, *e.g.* in Fig. 1, the question "when did we last visit the zoo?" needs to be answered by the "zoo" photos and "timestamp" of the photo together. The rich multimodal information may not always present in the existing VQA datasets.

To this end, this paper introduces the *MemexQA dataset*, the first publicly available dataset that contains real-world personal photo albums and questions about events captured in these photo sets. We crowdsource 20,563 questions and answers on 13,591 personal photos from 101 real Flickr users. Out of the 13K photos, about 5K photos are essential for answering the questions and we have mainly verified the answer on these 5K photos. These personal photos capture a wide variety of events of a user's life such as a trip to Japan, a wedding ceremony, a family party, etc. The annotator is instructed to ask the question about the main event/topic in all photos of a single user. The question is supposed to be useful in recalling the memory of these events. An answer includes a text answer as well as a few evidential photos that justify the answer. MemexQA is an interesting yet challenging task. Answering a MemexQA question is non-trivial even for adults. As shown in our experiments, it takes an adult more than a minute to answer a MemexQA question, which is *10x* longer than answering a VQA question. We consider MemexQA to be a multimodal AI task and may catalyze interesting real-world applications on the ever-increasing personal multimedia collection.

There are two challenges in this new task.

First, personal albums involve rich information. A user usually has multiple albums, as sequences of videos or images, ordered according to their time stamps. For every photo or video, the user may provide text annotations, tags, and other metadata. In this paper, we call such input as ***visual-text sequence*** data. Note that not all the photos and videos are annotated, which requires a robust method to leverage inconsistently available multimodal data.

The second challenge regards interpretable justifications in addition to direct answer based on sequence data. To help users with a lot of photos and videos, a natural requirement is to identify the supporting evidence for the answer. An example question as shown in Fig. 1, is "when did we last visit the zoo?" From the users' viewpoint, a good QA system should not only give a definite answer (e.g. April 11 2008), but also ground evidential images or text snippets in the input sequence to justify the reasoning process. The inspection process may be trivial for a single image but can take a significant amount of time to examine every image and the complete text words. We found that humans often need photo evidence to quickly verify the answer.

To address these two challenges, we propose the ***Focal Visual-Text Attention (FVTA)*** model for visual-text sequential data. Our model is motivated by the reasoning process of humans. In order to answer a question, a human would first quickly skim the input and then focus on a few, small temporal regions in the visual-text sequences to derive an answer. In fact, our statistics suggest that, on average, humans only need 1.5 images to answer a question after the skimming. Inspired by this process, FVTA first learns to localize relevant information within *a few, small, temporally consecutive regions* over the input sequences, and learns to infer an answer based on the cross-modal statistics pooled from these regions. FVTA proposes a novel kernel to compute the attention tensor that jointly models the latent information in three sources: 1) answer-signaling words in the question, 2) temporal correlation within a sequence, and 3) cross-modal interaction between the text and image. FVTA attention allows for collective reasoning by the attention kernel learned over a few, small, consecutive sub-sequences of text and image. It can also produce a list of evidential images/texts to justify the reasoning. To summarize, the contribution of this paper is threefold:

- We introduce a new multimodal QA task, MemexQA, as well as the first benchmark that contains questions about real-world personal photo albums.
- We propose a novel attention kernel for VQA on visual-text data. The proposed attention tensor can be used to localize evidential image and text snippets to explain the reasoning process.
- We empirically evaluate the performance of representative VQA models as well as the human performance on this new QA tasks, establishing the first experimental benchmark for future research to explore.

## 2 Related Work

***Consumer Photo Albums Understanding*** has been an important research topic in the multimedia community. An album usually contains a sequence of photos (some may also contain videos), where a personal photo is associated with rich metadata such as time, title, tags, or GPS information, and the sequence of photos are arranged in the temporal order and a coherent context. A number of methods have been developed to explore the photo understanding or annotation within the album context [7], [8], [9], [34], [62], especially on identifying persons and faces [30], [31], [33], [63]. In this work, we improve previous researches by considering personal photo albums in the ***Memex*** context. The term "Memex" was first posited by V. Bush in 1945 as an enlarged intimate supplement to an individual's memory. Bush envisioned the Memex as a device in which individuals would compress and store all of their information, mechanized so that it may be consulted with exceeding speed and flexibility [6]. The concept of the Memex influenced the development of early hypertext systems, eventually leading to the creation of the World Wide Web [12]. The proposed MemexQA focuses on the deep understanding of the multimedia contents of albums, by answering questions related to personal photos.

***VQA Datasets***. Our MemexQA work is partly motivated by the problem of Visual QA (VQA), which has received a large amount of attention and more than ten datasets have emerged to promote the research [3], [4], [14], [16], [20], [24], [25], [37], [38], [45], [50], [56], [57], [58], [60], [64], [65]. More recently, Kafle et al. [26] proposed a VQA dataset for bar chart understanding, which requires processing words and answers that are unique to a particular bar char. Liu et al. [35] proposed the inverse problem of visual question answering, iVQA, which was to generate a question that corresponds to a given image and answer pair. Agrawal et al. [1] proposed a new setting for VQA, VQA-CP, to overcome priors for visual question answering. VizWiz [17] collected visual questions from blind people and targeted to build systems that could assist blind people. Embodied Question Answering [11], [15], where an agent is spawned at at a random location in a 3D environment and answers questions by exploring the environment, was proposed to test a range of AI skills including language understanding and commonsense reasoning. Due to the difference in problem settings, these datasets are not directly applicable to our MemexQA research. Compared to existing VQA datasets,

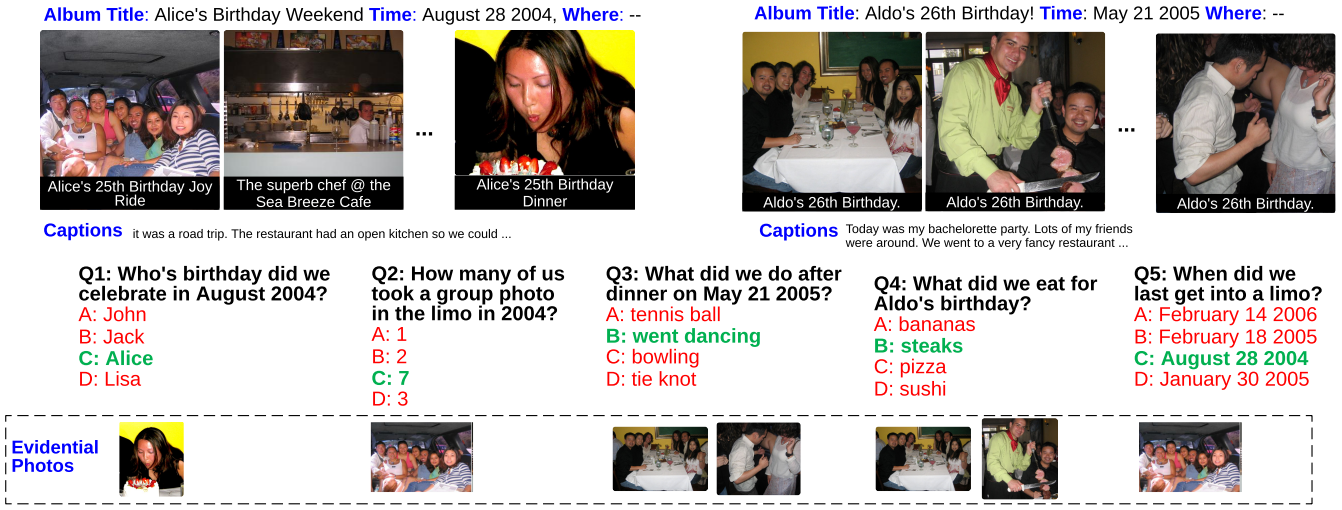



Fig. 2: Questions and four-choice answer in MemexQA. From the top to bottom are album metadata, photos from 2 albums, titles and captions, questions, answer choices, and evidential photos. The green choice denotes the correct answer.

the MemexQA dataset is distinguished by the following features. First, it is a goal-driven QA task over the photo sequence. By answering the question and showing relevant photos, our goal is to help users recover their memory about the asked event. Second, it provides ground-truth evidential photos used by human to answer the question. Third, MemexQA is a multimodal dataset. A big proportion of questions need to be answered using both the image and the text metadata.

***Attention-based QA methods.*** In text domain, machine comprehension is a major task and big datasets like CNN/DailyMail [18] and SquAD [44] have sparked many deep neural network models [18], [46], [47], [54]. In the machine comprehension task, systems are given a context paragraph of text to answer question about it. [18] proposed a dynamic attention mechanism where attention weights are updated dynamically based on the question and the context text plus the previous attention. [47] added memory network to compute attention weights and utilized multi-turn reasoning. [46] calculated a similarity matrix between each question word and each context word and used it to get bi-directional attention.

Many attention mechanisms have been proposed to allow the model to find important area in single image. [48] projected the question representation and the image region features into a common feature space to compute the attention weights for image region feature concatenated with question vector. [10] compares VQA models' attention with human attention. [57] computed a correlation matrix between each question word and each part of the image, and then utilized a two-hop process to attend to the correct area. The attended image feature is the input to the second hop attention. [36] also computed a correlation matrix, and added a layer to use it as a feature to compute the attention weights for image and question. [59] used a stacked dynamic attention where each layer's attention is computed using last layer's attended feature and the original image. [40] used similar dynamic attention but with memory network. More recenlty, bottom-up and top-down attention network [2] using object detection model was proposed to enable attention to be calculated at the level of objects and other salient image regions. Nguyen et al. proposed stacked dense symmetric co-attention [41] that formed a hierarchy for multi-step interaction between an image-question pair. Differential attention [42] was proposed to bridge the gap between human attention and VQA model attention. Recent state-of-the-art methods' attention mechanism has been focusing on different fusion strategies between multimodal representations [5], [13].

This work can be viewed as a novel attention model for multiple variable-length sequential inputs, to take into account not only the visual-text information but also the temporal dependency. Our work extends the previous studies of using attention model for Image QA [5], [10], [13], [36], [40], [48], [57], [59]. A key difference between our method and classical attention model lies in the fact we are modeling the correlation at every time step, across multiple sequences. Existing attention mechanisms for VQA mainly focus on attention within spatial regions of an image [65] or within a single sequence [20], and hence, may not fully exploit the multiple sequences and multiple time steps nature. As Fig. 6 shows, our attention is applied to a three-dimensional tensor, while the classic soft attention model is applied to a vector or matrix.

## 3 MemexQA Dataset

### 3.1 Data Collection

The data are annotated using Amazon Mechanical Turk (AMT), an online crowd-sourcing platform. Fig 3 shows the web interface of annotation. The annotation process can be viewed as 3 steps: QA collection, candidate answers generation, and QA cleanup.

***Questions & Answers (QA) Collection.*** We randomly select 101 real Flickr users from a public Flickr photo subset called SIND [19]. The photos are under Creative Commons BY-NC-ND 2.0 license. Each user has at least 4 albums. In total, the dataset consists of 20,563 question-answer pairs and of 13,591 personal photos from 101 real Flickr users, organized in 630 albums. During QA cleanup, 5,090 photos are shown to the annotators to verify. For each photo, we collect a variety of multimodal metadata which includes the timestamp, GPS location, photo title, tags, album information, and captions, in which the GPS, title, tag, album information, and captions can be missing for some photos.

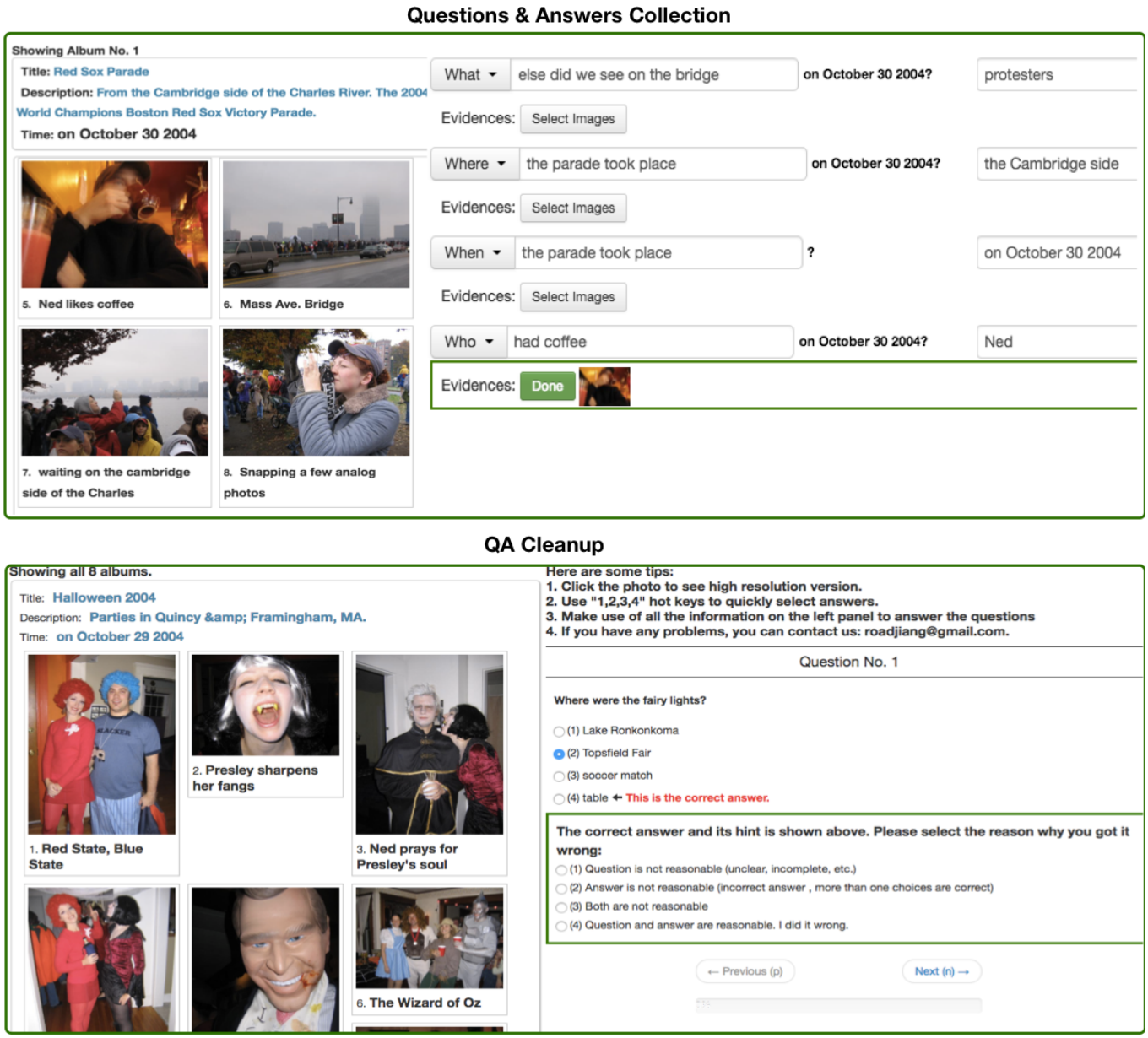



Fig. 3: Example of the annotation web interface. During questions and answers collection, annotators are given all the photos and metadata information of the assigned Flickr user on the left of the screen. On the right, annotators are asked to generate different types of questions and the corresponding answers. Meanwhile, they are also required to select evidence photos for each of the question. During QA cleanup stage, annotators are asked to select answers for the questions given all relevant information. If they select the wrong answers, they are asked to provide a reason why they make the mistakes, which will be used for the QA cleanup.

The online AMT workers are instructed to write important and objective questions. Also, the answers need to be concise about the main event or topic in the albums of a Flickr user. Each worker is asked to write questions as if they were his/her own personal photos. Specifically, one worker writes 4 questions/answers about a single album, and for each answer, he is required to label one or more evidential photos to justify the answers. Then for all the user's albums, the worker writes 12 questions and similarly provide evidential photos for each answer. Fig. 2 shows some examples questions, where the first three are questions within a single album whereas the rests are for two albums.

Motivated by [21], we focus on five types of questions: "what", "who", "where", "when" and "how many". As shown in the previous study, the query terms in these categories are estimated to account for more than 60% of Flickr's personal photo search traffic. We choose not to include the "show me" questions, as such questions should be addressed by a separate text-to-image/video module [23]. We acknowledge the assumption for collecting the QA data: we require all questions can be answered by any individual and not just by the owner of the album. Therefore, the information beyond the photos and the metadata should not be used in answering questions. Following this assumption, anyone can ask and answer questions about the event as long as they provide convincing evidential photos to justify the answer. This leads to a more objective evaluation protocol.

***Candidate Answers Generation.*** Following [65], we employ

TABLE 1: Comparison of representative VQA datasets and our new dataset called MemexQA.

| Datasets | Data source | Sequence input | specific goal-driven | Supporting/Grounding Evidence | Personal media collection | metadata time&GPS |
|---|---|---|---|---|---|---|
| DAQUAR [37] | Depth images | - | - | - | - | - |
| VQA [4], [16] | Web images (MSCOCO) | - | - | - | - | - |
| FM-IQA [14] & TDIUC [25] | Web images | - | - | - | - | - |
| FVQA [52] | Web images | - | - | supporting facts | - | - |
| Visual Genome [29], [65] | Web images (Flickr) | - | - | image pixels | - | - |
| VizWiz [17] | Personal photos | - | ✓ | - | ✓ | |
| MSRVTT-QA [56] | Web videos (MSRVTT) | ✓ | - | - | - | - |
| TGIF-QA [20] | Gif images | ✓ | - | - | - | - |
| MovieQA [50] | Movies | ✓ | ✓ | movie plots | - | - |
| Ours MemexQA | Personal photos | ✓ | ✓ | evidential photos | ✓ | ✓ |

both human workers and automatic methods to generate a pool of candidate answers as the multiple choice, which include four-choice and twenty-choice answers. For the "what" question, candidate answers are automatically generated based on the answers to similar questions in the MemexQA dataset as well as external datasets such as VQA [4] and Visual Genome [29]. For other types of questions, candidate answers are obtained by randomly selecting relevant user metadata. For example, for "when" questions, the candidate answers are the dates of the user's other photos. For twenty-choice answers, we balance the choice by selecting relevant candidate answers from all question types. The candidate answers are then inspected by annotators to ensure there is only one correct answer for each question.

***QA Cleanup.*** All questions and answers need to be unambiguous, objective and relevant to the event or topic of the photo. To control the quality, each photo album is independently annotated by at least three AMT workers. The QA along with candidate answers are verified by another three workers, where they are asked to select the correct answer from the provided multiple choices. The workers report unreasonable questions or answers when they find subjective questions, incorrect answers, or questions of more than one correct answers. We remove questions with more than two worker reports. Besides, we also manually screen the data and reject all QA pairs from low-quality AMT workers. As a result, the collected data are of decent quality. The sampled inter-human agreement is 0.9, which measures the percentage of the questions having the same answer cross different AMT workers. This number is comparable to existing representative VQA datasets [4], [65].

## 3.2 Data Characteristics

MemexQA is the first publicly available dataset composed of real-world personal photo albums and questions about events captured in these photo sets. With the focus on questioning photo collection, we believe that this new dataset nicely complements other VQA benchmarks (see Table 1 for a list of related VQA datasets and their differences with MemexQA) and that it would be an ideal benchmark for language and vision research on a real-world problem. Generally, compared to existing VQA datasets [3], [4], [14], [16], [20], [24], [25], [45], [50], [57], [58], [60], [65], the MemexQA dataset contains a few distinguishing characteristics. First, MemexQA defines a goal-driven VQA task over a user's personal photos, where the goal is that by answering questions, we help the users recover their memory in these photos. Personal photos often capture a wide variety of real-life events with great sentimental value, such as the marriage proposal, graduation ceremony, family gathering, birthday party, etc. Second, the MemexQA dataset contains a number of evidential photos for each question. For example, for Q4 in Fig. 2, the evidential photos show the birthday dinner and steak that help users quickly verify the answer "steak". Likewise, for Q5 the evidential photo shows the group photo taken on the limo and its timestamp, which justify the answer to the question "when did we last get into a limo?". Finally, MemexQA is a multimodal dataset. It encompasses rich information including time, GPS location, metadata, etc. The multimodal information makes it an ideal testbed for vision+language research. Fig. 2 also shows the diverse sources from which an answer can be derived, e.g. from title+time (Q1), image+time (Q2, Q3, Q5), image+time+title (Q4).

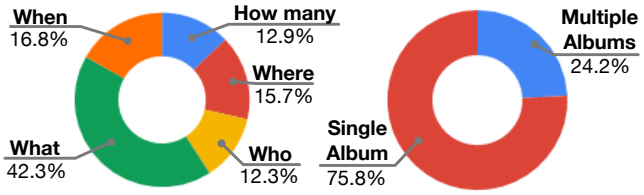



Fig. 4: Question distributions by question types and question albums/topics.

### 3.2.1 Question Types and Topics

There are 5 types of questions in this dataset. The statistics of question types are shown in Fig 4. Each question can be about either a single album/topic (Fig. 2 Q1-Q3) or across multiple albums/topics (Fig. 2 Q4 and Q5). The question distribution by relevant albums is also shown in Fig. 4.

### 3.2.2 Question Difficulty

The level of difficulty of this dataset can also be indicated in Table 2, where the percentage of answers directly found in text metadata is shown. With 32% on "What" questions and 46% overall, it suggests that to find the correct answers, the system has to reason and come up with words or phrases that do not exist in the metadata most of the time.

### 3.2.3 Limitations

Here we discuss the limitations of the dataset: 1) we assume all questions should be answerable by any individual and not just the owner of the photos. Therefore, information that is not captured by the photo or metadata should not be used in answering questions.

TABLE 2: Percentage of answers found in text metadata excluding answers from "How many" questions.

| Question Type | Answer Found in Text Metadata |
|---|---|
| What | 32.61% |
| When | 70.86% |
| Where | 48.36% |
| Who | 55.82% |
| Overall | 46.09% |

This may lead to lower recall yet a more objective approach for evaluation. 2) the scale of the dataset is smaller compared to the real-world setting. Each user only has 130 photos on average. This is due to the reason that MemexQA is very expensive to collect. In MemexQA, to write a question, AMT workers need to not only inspect dozens of photos but also consult the supporting multimodal information. As a result, we found that it takes on average 96±10 seconds for annotators to write a MemexQA question and 62±3 seconds to answer a question. It is about 10 times longer than writing or answering a question about a single image. Given those restrictions, we consider the MemexQA dataset to be reasonably large as the first benchmarks on which different methods for this novel task can be compared. As shown in Section 6 , the datasets are sufficiently large to train deep QA networks of reasonable performance.

### 3.3 Human Performance

We examine the human performance on MemexQA. We are interested in measuring 1) how well human can perform in the MemexQA task, 2) what is the contribution of each modality in helping users answer questions, and 3) how long does it take for humans to answer a MemexQA question.

We conduct a series of controlled experiments, which is evaluated with more than 150 human subjects (AMT workers). The workers are asked to select an answer from 4 choices given different information, which includes **Q**uestions, **A**nswers, **I**mages, and **M**etadata (titles, descriptions, timestamps and GPS (if any)). Table 3 summarizes the results. For example, Q+A+I indicates the human performance of choosing the correct answer by only looking at the question, answers, and images. As we see, humans manage to correctly guess 50% of the correct answers using common sense. With all information, the accuracy reaches 0.93, which is comparable to that on other VQA datasets (0.83 on VQA [4] and 0.97 on Visual7W [65]). The accuracy without images drops significantly, which indicates that the MemexQA task requires multimodal reasoning based on both vision and language. We record the time spent on each QA pair and found it takes on average 62 seconds for a human to answer a question, which is 10 times longer than answering a VQA question (about 5.5 seconds [65]). This suggests that an automatic system with exceeding speed and decent accuracy will provide great benefit to users.

TABLE 3: Human Performance on MemexQA. Q, A, I and M denote question, answer, image and metadata, respectively.

| Input | how many | what | when | where | who | overall |
|---|---|---|---|---|---|---|
| Q+A | 0.57 | 0.41 | 0.50 | 0.52 | 0.46 | 0.52 |
| Q+A+I | 0.93 | 0.73 | 0.90 | 0.85 | 0.76 | 0.86 |
| Q+A+M | 0.71 | 0.60 | 0.77 | 0.64 | 0.56 | 0.67 |
| Q+A+I+M | 0.94 | 0.87 | 0.96 | 0.96 | 0.86 | 0.93 |

## 4 Approach

### 4.1 Problem Formulation

We start the discussion by formally defining the problem. Let $Q = q_1, \cdots, q_M$ represent a question of $M$ words $Q \in \mathbb{Z}^M$, where each word is an integer index in the vocabulary. Define a context visual-text sequence of $T$ examples $\mathbf{X} = \mathbf{x}_1, \cdots, \mathbf{x}_T$, where for each example, $\mathbf{x}_t^{img}$ represents an image. $\mathbf{x}_t^{txt}$ is its corresponding text sentence, where its $i$-th word is indexed by $\mathbf{x}_{ti}^{txt}$. Following [4], [65], the answer to a question is an integer $y \in [1, L]$ over the answer vocabulary of size $L$. Given a collection of $n$ questions and their context sequences, we are interested in learning a model maximizing the following likelihood:

$$\underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(y_i | Q_i, \mathbf{X}_i; \Theta) \tag{1}$$

where $\Theta$ represents the model parameters. Given the visual-text sequence input $\mathbf{X}^{img}, \mathbf{X}^{txt}$, we obtain a good joint representation by attention model. With FVTA attention, the model takes into account of the sequential dependency in image or text sequence, respectively, and cross-modal visual-text correlations. Meanwhile, the computed attention weights over input sequences can be utilized to derive meaningful justifications.

### 4.2 Network Architecture

This subsection discusses our overall neural network architecture. As shown in Fig. 5, the proposed network consists of the following layers.

**Visual-Text Embedding** Every image or video frame is encoded with a pre-trained Convolutional Neural Network. Both word-level and character level embedding [28] are used to represent the word in text and question.

**Sequence Encoder** We use separate LSTM networks to encode visual and text sequences, respectively, to capture the temporal dependency within each individual sequence. The inputs to the LSTM units are image/text embedding produced by the previous layer. Let $d$ denote the size of the hidden state of the LSTM unit; the question $Q$ is represented as a matrix $\mathbf{Q}$ of concatenated bi-directional LSTM outputs at each step, i.e. $\mathbf{Q} \in \mathbb{R}^{2d \times M}$, where $M$ is the maximum length of the question. Likewise, The sequentially encoded text and images are represented by $\mathbf{H} \in \mathbb{R}^{2d \times T \times 2}$, where $T$ is the maximum length of the sequence.

**Focal Visual-Text Attention** The FVTA is a novel layer to implement the proposed attention mechanism. It represents a network layer that models the correlations between questions and multi-dimensional context and produces the summarized input to the final output layer, *i.e.*, $\tilde{\mathbf{h}} \in \mathbb{R}^{2d}$ and $\tilde{\mathbf{q}} \in \mathbb{R}^{2d}$. We will discuss FVTA in the next section.

**Output Layer** After summarizing the input using the FVTA attention, we use a feed-forward layer to obtain the answer candidate. For open-ended setting as described in Section 6, where there is no choice input, the task is to find the answer given the context and the question and the final probability across all possible answers is given by $\mathbf{p} = softmax(\mathbf{w_p^T}[\tilde{\mathbf{q}}; \tilde{\mathbf{h}}; \tilde{\mathbf{q}} \odot \tilde{\mathbf{h}}])$, where the operator $[\cdot;\cdot]$ represents the concatenation of two matrices or vectors along the last dimension. $\odot$ is the element-wise multiplication, $\mathbf{w_p}$ is the weight vector to learn and $\mathbf{p}$ is a vector of classification probability. In practice we find this simple equation works better than fully connected layer or straightforward concatenation. For multiple-choices questions, let $k$ denote the number of candidate answer choices, we utilize the bi-directional LSTM to encode

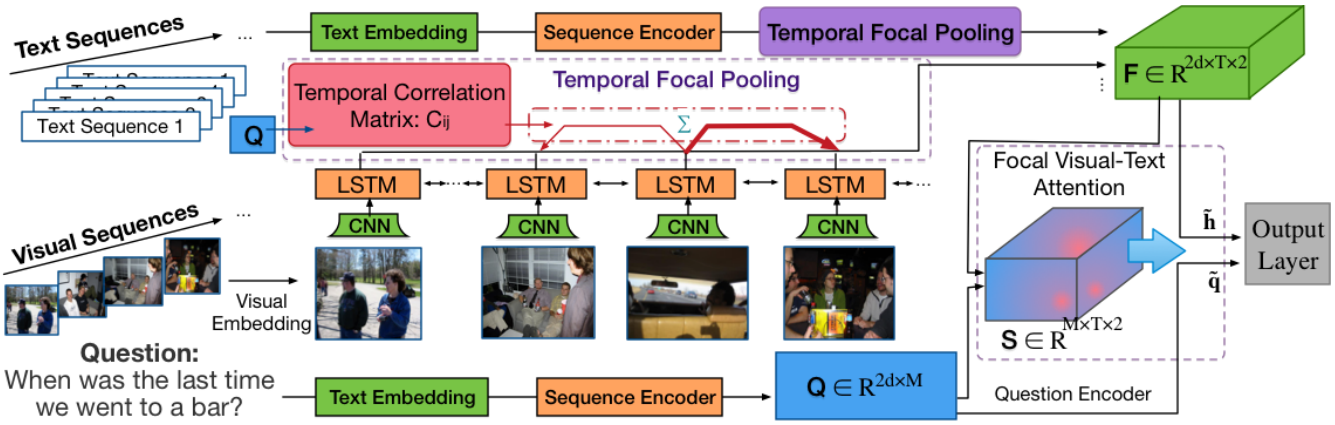



Fig. 5: An overview of Focal Visual-Text Attention (FVTA) model. For visual-text embedding, we use a pre-trained convolutional neural network to embed the photos and pre-trained word vectors to embed the words. We use a bi-directional LSTM as the sequence encoder. All hidden states from the question and the context are used to calculate the FVTA tensor. Based on the FVTA attention, both question and the context are summarized into single vectors for the output layer to produce final answer. Here the output layer is for open-ended setting while in multiple-choice setting, the text embedding of the answer choice is also used as the input, which is not shown in this figure.

each of the answer choice and use the last hidden state as the representation for answers $\mathbf{E} \in \mathbb{R}^{k\times 2d}$. We tile the context representation $\tilde{\mathbf{h}}$ and attended question representation, $k$ times into $\tilde{\mathbf{H}} \in \mathbb{R}^{k\times 2d}$ and $\tilde{\mathbf{Q}} \in \mathbb{R}^{k\times 2d}$ to compute the classification probability of $k$ choices:

$$\mathbf{p} = softmax(\mathbf{w_p^T}[\tilde{\mathbf{Q}}; \tilde{\mathbf{H}}; \mathbf{E}; \tilde{\mathbf{Q}} \odot \mathbf{E}; \tilde{\mathbf{H}} \odot \mathbf{E}]) \quad (2)$$

After obtaining the answer probability, the model can be trained end-to-end using cross-entropy loss function.

## 5 Focal Visual-Text Attention

This section discusses the details of FVTA model as the key module in our VQA system. We first introduce similarity metric between visual and text features, then discuss constructing the attention tensor that captures both intra-sequence dependency and inter-sequence interaction.

### 5.1 Similarity between visual and text features

To compute the similarity across different modalities, i.e. visual and text, we first encode every modality by the LSTM networks with the same size of hidden states. Then we measure the differences between these hidden state variables. Following the study in text sequence matching [53], we aggregate both the cosine similarity and Euclidean distance to compare the features. Moreover, we choose to keep the vector information instead of summing up after the operation. The vector representation can be used as the input of a learning model, whose inner product represents the similarity between these features. More specifically, we use the following equation to compute the similarity representation between two hidden state vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. The result is a vector of twice the hidden size:

$$\mathbf{s}(\mathbf{v}_1, \mathbf{v}_2) = [(\mathbf{v_1} \odot \mathbf{v}_2); (\mathbf{v}_1 - \mathbf{v}_2) \odot (\mathbf{v}_1 - \mathbf{v}_2)]. \quad (3)$$

### 5.2 Intra-sequence temporal dependency

Our visual-text attention layer is designed to let the model select related visual-text region or timestep based on each word of the question. Such fine-grained attention is in general nontrivial to learn. Meanwhile, most answers for visual-text sequence inputs may be constrained and restricted in a short temporal period. We learn such localized representation, called focal context representation, to emphasize relevant context states based on the question.

First, we introduce a temporal *correlation matrix*, $\mathbf{C} \in \mathbb{R}^{T\times T}$, a symmetric matrix where each entry $c_{ij}$ measures the correlation between context's the $i$-th step and the $j$-th step for a question. Let $\mathbf{h}_i = \mathbf{H}_{:i:} \in \mathbb{R}^{2d\times 2}$ denote the visual/text representation for the $i$-th timestep in $\mathbf{H}$. For notation convenience, : is a slicing operator to extracts all elements from a dimension. For example, $\mathbf{h}_{i1} = \mathbf{H}_{:i1}$ represents the vector representation of the $i$-th timestep of the visual sequence. Here we denote the last index 1 for visual and 2 for textual modality. Each entry $\mathbf{C}_{ij}$ ($\forall i, j \in [1, T]$) is then calculated by:

$$\mathbf{C}_{ij} = \tanh \sum_{k=1}^{2} \mathbf{w}_c^\top (\mathbf{w}_h^\top \mathbf{s}(\mathbf{h}_{ik}, \mathbf{h}_{jk}) + \mathbf{Q}_{:M}) \quad (4)$$

where $\mathbf{w}_c \in \mathbb{R}^{2d\times 1}$ and $\mathbf{w}_h \in \mathbb{R}^{4d\times 2d}$are parameters to learn. The temporal correlation matrix captures the temporal dependency of question, image and text sequence.

To allow the model to capture the context between timesteps based on the question, we introduce temporal focal pooling to connect neighboring time hidden states if they are related to the question. For example, it can capture the relevance between the moment "dinner" and the moment later, "Went dancing", given the question "What did we do after the dinner on Ben's birthday?". Formally, given the time correlation matrix $\mathbf{C}$ and the context representation $\mathbf{H}$, we introduce a *temporal focal pooling function* $g$ to obtain the focal representation $\mathbf{F} \in \mathbb{R}^{2d\times T\times 2}$. Each vector entry $\mathbf{F}_{:tk}$ ($\forall t \in [1, T], \forall k \in [1, 2]$) in $\mathbf{F}$ is calculated by:

$$\begin{aligned}\mathbf{F}_{:tk} &= g(\mathbf{H}; \mathbf{C}, t, k) \\ &= \sum_{s=1}^{T} \mathbb{1}[s \in [t-c, t+c]]\mathbf{C}_{st}\mathbf{h}_{sk},\end{aligned} \quad (5)$$

where $\mathbf{F}_{:tk}$ is the focal context representation at $t$-th timestep for visual ($k = 1$) or text ($k = 2$). $\mathbb{1}$ is the indicator function. $c$ stands for the size of the temporal window that is a hyper-parameter. We constrain the model to focus on a few small temporal context

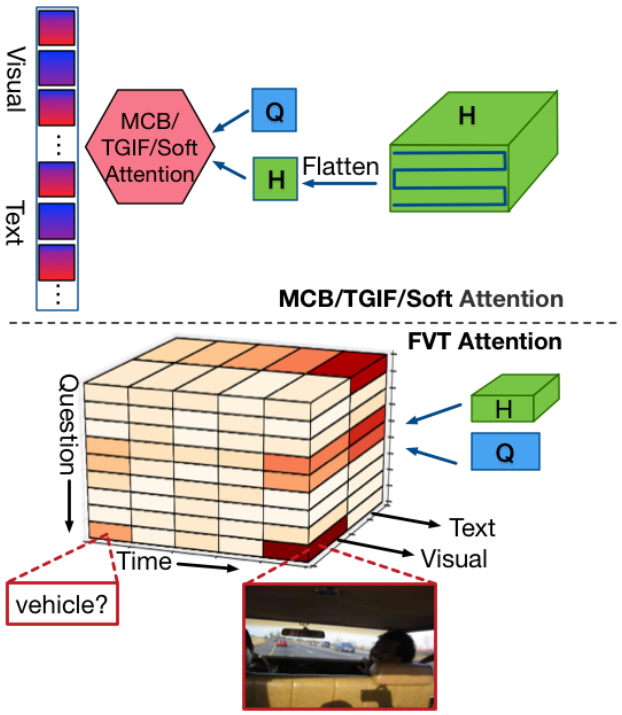



Fig. 6: Comparison of our FVTA and classical VQA attention mechanism. FVTA considers both visual-text intra-sequence correlations and cross sequence interaction, and focuses on a few, small regions. In FVTA, the multi-modal feature representation in the sequence data is preserved without losing information.

windows of size $2c+1$. We design this function based on the intuition that most answers are constrained in a short temporal period and its efficacy has been proven in the ablation studies.

### 5.3 Cross Sequence Interaction

In the previous section, we describe how we construct a sequential context representation within each modality. In this section, we introduce the attention mechanism to capture the important correlation between visual and textual sequences. We apply such attention over the focal context representation to summarize important information for answering the question.

Firstly, we obtain the attention weights based on a tensor $\mathbf{S}$ between each word of the question and each timestep of the visual-text sequences. We computes a *kernel tensor*, $\mathbf{S} \in \mathbb{R}^{M\times T\times 2}$, between the input question and the focal context representation $\mathbf{F}$, where each entry in the kernel $s_{mtk}$ models the correlation between the $m$-th word in question and at $t$-th timestep over the modal $k$ (images or text words). Let $\mathbf{v}_{tk}$ denote the focal context representation $\mathbf{F}_{:tk}$ at $t$-th timestep for visual or text. Each entry $s_{mtk}$ in $\mathbf{S}$ is calculated by:

$$\begin{aligned} s_{mtk} &= \kappa(\mathbf{F}_{:tk}, \mathbf{Q}_{:m}) = \kappa(\mathbf{v}_{tk}, \mathbf{q}) \\ &= \tanh(\mathbf{w}_s^\top \mathbf{s}(\mathbf{v}_{tk}, \mathbf{q}) + \mathbf{b}_s) \end{aligned} \tag{6}$$

where $\kappa$ is a function to compute the correlation between question and context, $\mathbf{w}_s \in \mathbb{R}^{4d\times 1}$ is the learned weights and $\mathbf{b}_s$ is the bias term. $\mathbf{s}$ is the mapping defined in (3). As explained for Eq. (4), we use such similarity representations since they capture both the cosine similarity and Euclidean distance information.

Based on attention weight tensor $\mathbf{S}$, we obtain the visual-text sequence attention matrix $\mathbf{A} \in \mathbb{R}^{T\times 2}$ by $\mathbf{A} = softmax(\max_{i=1}^{M}(\mathbf{S}_{i::}))$ and the visual-text attention vector $\mathbf{B} \in \mathbb{R}^2$ by $\mathbf{B} = softmax(\max_{i=1}^{T}\max_{j=1}^{M}(\mathbf{S}_{ji:}))$, where the softmax operation is applied to the first dimension. The maximum function $\max_i$ is used reduce the first dimension of the high-dimensional tensor. Then the attended context vector is given by:

$$\tilde{\mathbf{h}} = \sum_{k=1}^{2} \mathbf{B}_k \sum_{t=1}^{T} \mathbf{A}_{tk}\mathbf{F}_{:tk} \in \mathbb{R}^{2d} \tag{7}$$

Similarly, we compute the question attention $\mathbf{D} \in \mathbb{R}^M$ by $\mathbf{D} = softmax(\max_{i=1}^{T}\max_{j=1}^{2}(\mathbf{S}_{:ij}))$ and the summarized question vector is given by:

$$\tilde{\mathbf{q}} = \sum_{m=1}^{M} \mathbf{D}_m \mathbf{Q}_{:m} \in \mathbb{R}^{2d} \tag{8}$$

Algorithm 1 summarizes the steps to compute the proposed FVTA attention. To obtain a final context representation, we first summarize the focal context representation separately for visual sequence and text sequence, emphasizing the most important information using the intra-sequence attention. Then, we obtain the final representation by summing the sequence vector representation based on the inter-sequence importance. Fig. 6 illustrates the difference between FVTA attention tensor and one-dimensional soft attention vector. Both mechanisms compute the attention but FVTA considers both visual-text intra-sequence correlations and cross sequence interaction.

**Algorithm 1:** Computation of Focal Visual-Text Attention.

**input :** Input visual-text sequence $\mathbf{X}$, Question $Q$
**output:** The FVTA vector $\tilde{\mathbf{h}}$

1. Encode $\mathbf{X}$ into $\mathbf{H}$ by the visual-text embedding and sequence encoder in Sec. 4.2;
2. Encode Q into $\mathbf{Q}$ by the question encoder;
3. Compute $\mathbf{C}$ by Eq. (4) `// temporal correlation`
4. Compute $\mathbf{F}$ by Eq. (5) `// intra-sequence dependency`
5. Compute $\mathbf{S}$ by Eq. (6) `// cross-sequence interaction`
6. Reduce $\mathbf{F}$ with $\mathbf{S}$ to the FVTA $\tilde{\mathbf{h}}$ by Eq. (7);
7. **return** $\tilde{\mathbf{h}}$;

## 6 Experiments

### 6.1 Setup

This section evaluates the MemexQA task: given a question and a sequence of personal photos, the goal is to find an answer and a few grounding photos that support the answer. As other VQA datasets do not satisfy our problem setting, we choose the MemexQA dataset as the main testbed. We divide the training and test split with 85%, 15% of the total QA pairs respectively, and further sample 20% of the training pairs as the validation set. We use version 1.1 of the dataset and the 5k photo collection as described in Section 3 is used as context input. The dataset and models are released at https://memexqa.cs.cmu.edu.

**Evaluation Settings.** We evaluate a model in terms of *two input settings*, answering questions with or without choices input (***Multiple-choice Questions*** vs. ***Open-ended Questions***). Under each input setting, we evaluate the models in terms of the precision of selecting the correct answer out of ***4 or 20 candidate choices***. For the multiple-choice question setting, following the evaluation protocol in the VQA challenge [13], we embed answer choices as input to the model in the multiple-choice setting and the model will output the probabilities over the given candidate answers. For the open-ended question setting, no answer embedding is used

| Method | how many (11.8%) | what (41.9%) | when (16.2%) | where (17.2%) | who (12.9%) | overall |
|---|---|---|---|---|---|---|
| Bag-of-words + CNN features w/ SVM | - / 0.251 | - / 0.224 | - / 0.229 | - / 0.243 | - / 0.266 | - / 0.237 |
| Bag-of-words + Semantic features w/ SVM | - / 0.270 | - / 0.243 | - / 0.260 | - / 0.263 | - / 0.246 | - / 0.253 |
| Multi-Layer Perceptron | 0.520 / 0.714 | 0.284 / 0.478 | 0.152 / 0.297 | 0.151 / 0.229 | 0.051 / 0.266 | 0.238 / 0.406 |
| Embedding + LSTM | 0.759 / 0.771 | 0.457 / 0.602 | 0.269 / 0.356 | 0.367 / 0.405 | 0.323 / 0.369 | 0.429 / 0.518 |
| Embedding + LSTM + Concat | 0.773 / **0.804** | 0.489 / 0.662 | 0.277 / 0.366 | 0.396 / 0.465 | 0.371 / 0.426 | 0.457 / 0.567 |
| DMN+ [55] | **0.783** / 0.790 | 0.559 / 0.663 | 0.326 / 0.389 | 0.472 / 0.539 | 0.431 / 0.440 | 0.517 / 0.584 |
| Multimodal Compact Bilinear Pooling [13] | 0.761 / 0.800 | 0.558 / 0.681 | 0.319 / 0.377 | 0.490 / 0.571 | 0.473 / 0.545 | 0.521 / 0.609 |
| Bi-directional Attention Flow [46] | 0.766 / 0.790 | 0.525 / 0.689 | 0.283 / 0.356 | 0.484 / 0.567 | 0.415 / 0.468 | 0.493 / 0.598 |
| Soft Attention | 0.776 / 0.795 | 0.546 / 0.697 | 0.311 / 0.346 | 0.523 / 0.604 | 0.459 / 0.582 | 0.520 / 0.621 |
| TGIF Temporal Attention [20] | 0.764 / 0.761 | 0.532 / 0.700 | 0.271 / **0.522** | 0.434 / 0.582 | 0.382 / 0.477 | 0.481 / 0.630 |
| FVTA | 0.728 / 0.761 | **0.628** / **0.714** | **0.354** / 0.476 | **0.607** / **0.676** | **0.611** / **0.668** | **0.590** / **0.669** |

TABLE 4: Comparison of different methods on MemexQA multiple-choice setting by question type. The first five methods do not use the attention mechanism. We show both 20 and 4 choice evaluation in XXX/YYY, respectively. - are infeasible due to the high dimensionality of input features.

and the model will output probabilities over all possible answers. Since in MemexQA dataset there is only one answer per question, we take the argmax over the multiple choice answers at test time for evaluating open-ended questions.

**Evaluation Metrics.** In addition to the accuracy, we also measure the model's capability of finding correct grounding photos for answering questions by looking at whether the ground truth evidence photos are in the top-$k$ outputted grounding photo (***HIT@k***). An ideal model would have a high accuracy as well as a decent HIT@k.

**Baseline Methods.** A large proportion of the existing solutions is to project image or videos into an embedding space, and train a classification model using these embeddings. We implement the following methods as baselines: ***Bag-of-words + CNN features w/ SVM*** is a naive baseline where textual information is represented by bag-of-words feature and concatenated with Convolutional Neural Network features for photos, which is used for classification with Support Vector Machine. ***Bag-of-words + Semantic features w/ SVM*** utilizes high-level semantic feature for photos instead of CNN features. They are extracted using Google Cloud Image API. ***Multi-Layer Perceptron*** uses feed-forward layers to extract visual-text features into the same dimension and predicts the final answers with one fully connected layer using the concatenated image, question and metadata features. ***Embedding + LSTM*** utilizes word embeddings and character embeddings, along with the same visual embeddings used in FVTA. Embeddings are encoded by LSTM and averaged to get the final context representation. ***Embedding + LSTM + Concat*** concatenates the last LSTM output from different modalities to produce the final output. On the other hand, we compare the proposed model to a rich collection of VQA attention models: ***Classic Soft Attention*** uses classic one dimensional question-to-context attention to summarize context for question answering. A correlation matrix between each question word and context is used to compute the attention as in [46], [57]. ***DMN+*** is the improved dynamic memory networks [55], which is one of the representative architectures that achieve good performance on the VQA Task. We implement the DMN+ network with each sentence and each photo representation used in our proposed network as supporting facts input. ***Multimodal Compact Bilinear Pooling (MCB)*** [13] is a recent competitive method on VQA [4] dataset. The spatial attention in the original model is directly used on the sequential images input. The hyperparameters including the output dimension of MCB are selected based on the validation results. ***Bi-directional Attention Flow (BiDAF)*** implements the single-modal attention flow model [46] overall concatenated context representations with embeddings as in FVTA network. ***TGIF Temporal Attention*** [20] is a recently proposed spatial-temporal reasoning network on sequential animated image QA. Since other baseline methods do not use spatial attention, we compare the TGIF network with temporal attention only. TGIF temporal attention uses a simple MLP to compute the attention and only the last hidden state of the question is considered. We compute the attention following [20] and use the same output layer in our method.

**Implementation Details.** In MemexQA dataset, each question is asked to a sequence of photos organized in albums. A photo might have 5 types of textual metadata, including the *album title*, *album descriptions*, *GPS Locations*, *timestamp* and a *title*. We use $N$ to denote the maximum number of albums, $K$ for the maximum number of photos in an album and $V$ for the maximum words. For album-level textual sequences like album titles and descriptions, the $K$ dimension only has one item and others are zero-padded. We also use zeros to pad those positions with no word/image. We encode GPS locations using words. The photos and their corresponding metadata form the visual-text sequences. All questions, textual context and answers are tokenized using the Stanford word tokenizer. We use pre-trained GloVe word embeddings [43], which is fixed during training. For image/video embedding, we extract fixed-size features using the pre-trained CNN model, Inception-ResNet v2 [49], by concatenating the pool5 layer and classification layer's output before softmax. We then use a linear transformation to compress the image feature into 100 dimensional. Then a bi-directional LSTM is used for each modality to obtain contextual representations. Given a hidden state size of $d$, we concatenate the output of both directions of the LSTM and get a question matrix $\mathbf{Q} \in \mathbb{R}^{2d \times M}$ and context tensor $\mathbf{H} \in \mathbb{R}^{2d \times V \times K \times N \times 6}$ for all media documents. We reshape the context tensor into $\mathbf{H} \in \mathbb{R}^{2d \times T \times 6}$. We select the best hyper-parameters based on performance on the validation set except for the focal pooling window size, which is set to 3. We use cross-entropy loss for both the open-ended setting and multiple-choice setting. We use the AdaDelta [61] optimizer and an initial learning rate of 0.5 to train for 200 epochs with a dropout rate of 0.3. We find L2 weight regularization not useful therefore we don't use any. In the multiple-choice question answering setting, we also use character embedding concatenated with the word embedding across all baselines and the LSTM hidden size is set to 50. In the open-ended setting, the LSTM hidden size is 128 and we do not use temporal focal pooling for the FVTA model so that it has a similar number of parameters compared to TGIF and Soft

| Method | how many (11.8%) | what (41.9%) | when (16.2%) | where (17.2%) | who (12.9%) | overall |
|---|---|---|---|---|---|---|
| Bag-of-words + CNN features w/ SVM | 0.771 / 0.776 | 0.121 / 0.292 | 0.099 / 0.192 | 0.086 / 0.342 | 0.029 / 0.343 | 0.177 / 0.348 |
| Bag-of-words + Semantic features w/ SVM | 0.764 / 0.778 | 0.118 / 0.290 | 0.061 / 0.190 | 0.067 / 0.357 | 0.022 / 0.360 | 0.164 / 0.352 |
| Multi-Layer Perceptron | 0.757 / 0.766 | 0.084 / 0.249 | 0.080 / 0.215 | 0.046 / 0.314 | 0.013 / 0.295 | 0.147 / 0.322 |
| Embedding + LSTM | 0.754 / 0.757 | 0.166 / 0.309 | 0.272 / 0.305 | 0.199 / 0.359 | 0.191 / 0.330 | 0.262 / 0.372 |
| Embedding + LSTM + Concat | 0.723 / 0.728 | 0.192 / 0.328 | 0.314 / 0.326 | 0.202 / 0.365 | 0.215 / 0.345 | 0.279 / 0.384 |
| DMN+ [55] | 0.752 / 0.754 | 0.218 / 0.335 | 0.264 / 0.267 | **0.257** / 0.367 | **0.295** / 0.330 | 0.305 / 0.378 |
| Multimodal Compact Bilinear Pooling [13] | 0.740 / 0.742 | 0.208 / 0.411 | 0.387 / 0.391 | 0.206 / 0.490 | 0.244 / 0.492 | 0.304 / 0.471 |
| Bi-directional Attention Flow [46] | 0.711 / 0.730 | 0.252 / 0.411 | 0.297 / 0.305 | 0.240 / 0.444 | 0.257 / 0.424 | 0.312 / 0.439 |
| Soft Attention | 0.728 / 0.733 | 0.237 / 0.402 | 0.307 / 0.309 | 0.229 / 0.438 | 0.233 / 0.462 | 0.305 / 0.440 |
| TGIF Temporal Attention [20] | **0.780** / **0.783** | 0.223 / 0.346 | 0.326 / 0.339 | 0.220 / 0.365 | 0.290 / 0.369 | 0.314 / 0.403 |
| FVTA | 0.740 / 0.759 | **0.303** / **0.487** | **0.412** / **0.431** | 0.227 / **0.533** | 0.288 / **0.552** | **0.357** / **0.526** |

TABLE 5: Comparison of different methods on MemexQA open-ended setting by question type. The first five methods do not use the attention mechanism. We show both 20 and 4 choice evaluation in XXX/YYY, respectively.

Attention. It takes about 48 hours to train the FVTA model with temporal focal pooling and about 20 hours without on a Nvidia TITAN X GPU.

## 6.2 Baseline Comparison

### 6.2.1 Multiple-choice Questions

In the multiple-choice setting, we embed the candidate answers as input to the model. Table 4 compare the accuracy on the MemexQA with multiple-choice input. We evaluate all the models with both 4 and 20 answer candidate input for comprehensive comparison. As we see, the proposed method consistently outperforms the baseline methods and achieves the state-of-the-art accuracy on this dataset. The first 5 methods in the table show the performance of embedding methods without any attention. Although embedding methods are relatively simple to implement, their performance is much lower than the proposed FVTA model. The experiment results advocate the attention model among images and image sequences. Compare to previous attention models, our FVTA network significantly outperforms other methods, which proves the efficacy of the proposed method.

### 6.2.2 Open-ended Questions

We also compare FVTA to these methods under the harder open-ended question answering setting, where no answer candidate is provided during training and the model will output probabilities over all possible answers. During testing, we evaluate each model with both the 4 and 20 choice metrics where we select the highest probability answer output among the candidate choices. The result is shown in Table 5. As we see, the proposed method consistently outperforms the baseline methods overall again under the open-ended setting. Our approach also achieves the best performance across most of the question types and metrics.

### 6.2.3 Grounding Quality Comparison

Since the overall accuracy is not perfect, users need grounding photos to verify the answer. The attention mechanism outputs an attention value to each photo, which reflects the importance of the photo used in answering the question. We rank the photos using the attention values as the grounding photos. To evaluate the quality of these grounding photos, we compute the correlation (HIT@k) between the ground-truth evidential photos and the top-$k$ outputted grounding photos. A perfect correlation means the model derives the answer using the exact same photo used by human annotators. An ideal VQA model should not only enjoy a high accuracy in answering a question (Table 5, 4) but also can find images that are highly correlated to the ground-truth evidence photos. We compare the grounding quality among the multiple-choice setting models. Table 6 lists the accuracy to examine whether a model puts focus on the correct photos. FVTA outperforms other attention models on finding the relevant photos for the question. The results show that the proposed attention can capture salient information for answering the question. For qualitative comparison, we select some representative questions and show both the answer and the retrieved top images based on the attention weights in Fig. 9. As shown in the first example, the system has to find the correct photo and visually identify the object to answer the question "what did the daughter eat while her dad was watching during the trip in June 2010?". FVTA attention puts a high weight on the correct photo of the girl eating a corn, which leads to correctly answering the question. Whereas for soft attention, the one-dimensional attention network outputs the wrong image and gets the wrong answer. This example shows the advantage of FVTA modeling the correlation at every time step, across visual-text sequences over the traditional dimensional attention. Overall, FVTA not only outputs the correct answers but also gives the correct justifications, which is very useful in real life application where users would want to verify the system's results. While MCB and TGIF attention gets some of the answers correct, they output the wrong image.

| | HIT@1 | HIT@3 | mAP |
|---|---|---|---|
| MCB | 11.98% | 30.54% | 0.269±0.005 |
| BiDAF | 6.36% | 19.50% | 0.203±0.004 |
| Soft Attention | 1.16% | 12.60% | 0.168±0.002 |
| TGIF Temporal | 13.28% | 32.83% | 0.289±0.005 |
| FVTA | **15.48**% | **35.66**% | **0.312±0.005** |

TABLE 6: The quality comparison of the learned FVTA and classic attention. We compare the image of the highest activation in a leaned attention to the ground truth evidence photos which human used to answer the question. HIT@1 means the rate of the top attended images being found in the ground truth evidence photos. AP is computed on the photo ranked by their attention activation.

### 6.2.4 Model Complexity Comparison

To investigate the differences of model complexity, we compare the number of float-point operations and parameters between different methods. As shown in 7, where we plot the number of parameters for different models against their accuracy in the open-ended setting evaluation. As we see, FVTA achieves the best accuracy with slightly more parameters than most methods, and requires less parameters than the popular MCB method on

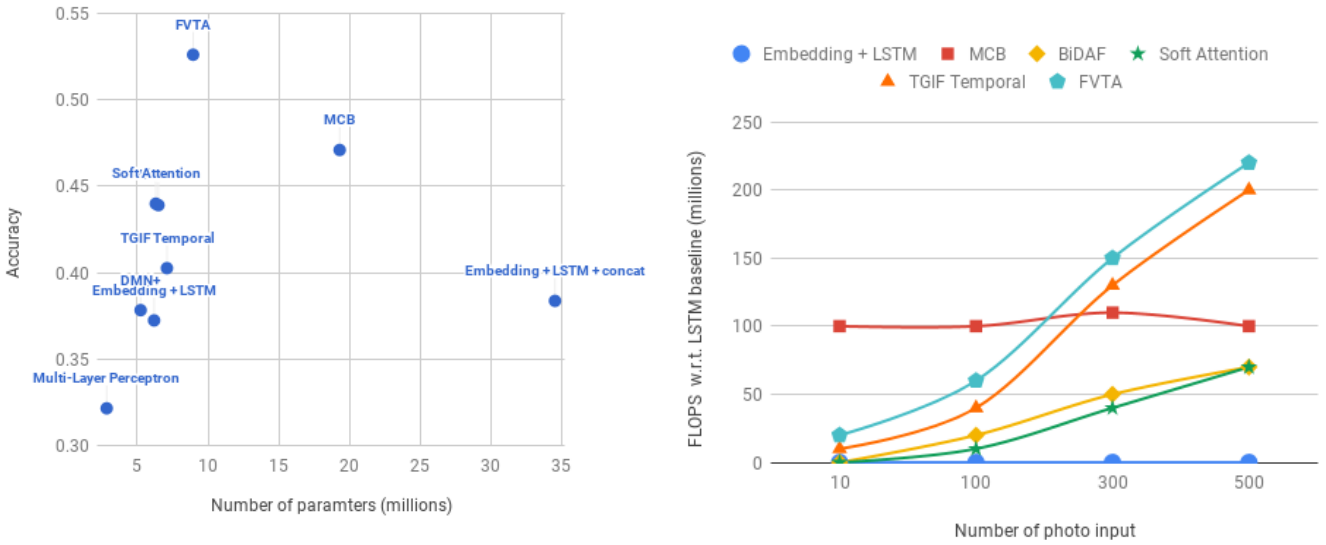



Fig. 7: Accuracy versus number of parameters (left) and computational demand versus the number of photo input (right) across different models. Computational demand is measured in the number of floating-point operation (FLOPS) to process the given number of photo input. The FLOPS numbers for other methods are in addition to the LSTM baseline's FLOPS. We only compare LSTM-based methods. Embedding + LSTM + concat has about 3x more FLOPS in general and thus omitted.


Fig. 8: Example failure cases of FVTA model on MemexQA dataset. Our model finds the correct grounding photos in both cases but fails to choose the correct answers due to the limitations discussed in Section 6.3.1.

VQA dataset. To compare computational demand, we compare the number of floating-point operations (FLOPS) across different methods. To have a fair comparison, we only compare methods that are based on LSTM. DMN+ uses a modified GRU which results in much fewer FLOPS. We subtract all other method's FLOPS with the baseline LSTM method and show the relative numbers. As we see, FVTA requires slightly more computation than TGIF as the number of photo input increases. In future work, we will look into reducing the computation demand for FVTA which is especially important for long sequence input.

### 6.3 Ablation Study

Table 7 shows the performance of FVTA mechanism and its ablations on the MemexQA dataset. We evaluate them under the multiple-choice question setting (except the last row) and with both 4 and 20 candidate answers. Firstly, we conduct a baseline experiment with only question inputs. LSTM is used to encode the question and the last hidden states are used directly to predict the answers. To evaluate the FVTA attention mechanism, we first replace our kernel tensor with simple cosine similarity function. Results show that standard cosine similarity is inferior to our similarity function on both metrics. For ablating intra-sequence dependency, we use the representations from the last timestep of each context document. For ablating cross sequence interaction, we average all attended context representation from different modalities to get the final context vector. Both aspects of correlation of the FVTA attention tensor contribute towards the model's performance, while intra-sequence dependency shows more importance in this experiment. We have also trained our model without temporal focal pooling, which means we use the context hidden states $\mathbf{H}$ to replace the use of focal context representation $\mathbf{F}$. The result shows that temporal focal pooling contributes to the model accuracy. We compare the effectiveness of context-aware question attention by removing the question attention and use the last timestep of the LSTM output from the question as the question representation. It shows the question attention provides slight improvement. Finally, we train FVTA with photo input only and text input only to see the contribution of visual and text information. Both results show that visual information serves a more important role but both visual and text information have significant influence. The result for the text-only model is good but it is perhaps not surprising due to the language bias in the questions and answers of the dataset, which is not uncommon in VQA dataset [4] and in Visual7W [65]. This also suggests significant rooms of improvement with visual information. In the last row of Table 7, we also conduct the photo-only model experiment under the open-ended question setting. Comparing the same model under multiple-choice setting, the relative performance drops are similar, suggesting the effect of visual-text information is similar despite different question settings.

#### 6.3.1 *Limitations & Error Analysis*

In this section, we investigate the limitations of the proposed method and conduct error analysis on the multiple-choice setting model.

| Ablations | 4-C | Δ | 20-C | Δ |
|---|---|---|---|---|
| Question-only | 0.405 | -26.3% | 0.213 | -37.6% |
| FVTA w/ Cosine Similarity | 0.619 | -4.9% | 0.499 | -9.0% |
| FVTA w/o Intra-seq | 0.569 | -10.0% | 0.507 | -8.2% |
| FVTA w/o Cross-seq | 0.604 | -6.5% | 0.566 | -2.9% |
| FVTA w/o Focal Pooling | 0.613 | -5.5% | 0.561 | -2.3% |
| FVTA w/o Question Attention | 0.629 | -4.0% | 0.553 | -3.7% |
| FVTA w/o Photos | 0.577 | -9.1% | 0.477 | -11.2% |
| FVTA w/o Metadata | 0.603 | -6.6% | 0.511 | -7.9% |
| FVTA w/o Metadata (Open-Ended) | 0.481 | -4.5% | 0.313 | -4.4% |

TABLE 7: Ablation studies of the proposed FVTA method on the MemexQA dataset. The last column shows the performance drop.

**Scalability.** As shown in the FLOP analysis in Fig 7, the gap between our model's computation demand and the baseline LSTM's grows linearly as the number of photo input increases. In the real-world scenario, personal photos are usually in the scale of thousands, which may render our method infeasible. In future work, additional methods to reduce the attention computation complexity over long sequences should be considered.

**Lack of Spatial Reasoning.** As shown in Fig. 8 on the left, even though our model selects the correct grounding photo, it fails to choose the correct answer due to lack of spatial reasoning abilities (the color of the pillow is "red" but the entire photo is mainly made of "red and yellow"). Although it is straightforward to add spatial attention to the current model, the problem of scalability will arise as directly adding spatial attention will add an order of magnitude of computation.

**Lack of Common Scene Understanding.** As shown in Fig. 8 on the right, the model finds the correct grounding snippet but is unable to choose the correct answer. It is clear that the model lacks common scene understanding capabilities and it seems to find the answer based on some language prior learned from the training set. One future direction could be to take advantage of low-level tasks like single-image Visual Question Answering and scene classification to incorporate common knowledge such as scene understanding.

### 6.4 MovieQA Experiments

To validate the efficacy of the proposed method, we also conduct experiments on the MovieQA dataset [50].

**Dataset** The MovieQA dataset consists of 140 movies and 6,462 multiple choice QA pair. Each QA pair contains five answer choices with only one correct answer. Systems are required to answer the questions given a number of movie clips from the same movie and the corresponding subtitles. More details of the dataset can be viewed in [50].

**Implementation Details** In the MovieQA dataset, each QA is given a set of $N$ movie clips of the same movie, and each clip comes with subtitles. We implement FVTA network for MovieQA task with modality number of 2 (video & text). We set the maximum number of movie clips per question to $N = 20$, the maximum number of frames to consider to $F = 10$, the maximum number of subtitle sentences in a clip to $K = 100$ and the maximum words to $V = 10$. Visual and text sequences are encoded in the same way as in the MemexQA experiment. We use the AdaDelta [61] optimizer with a minibatch of 16 and an initial learning rate of 0.5 to trained for 300 epochs. A dropout rate is set at 0.2 during training. The official training/validation/test split is used in our experiments.

**Experimental Results** We compare FVTA with recent results on MovieQA dataset, including End-to-End Memory Network (MemN2N) [51], Deep Embedded Memory Network (DEMN) [27], and Read-Write Memory Network (RWMN) [39]. Table 8 shows the detailed comparison of MovieQA results using both videos and subtitles. FVTA model outperforms all baseline methods and achieves comparable performance to the state-of-the-art result [4] on the MovieQA test server. Notably, RWMN [39] is a very recent work that uses memory net to cache sequential input, with a high capacity and flexibility due to the read and write networks. Our accuracy is 0.410 (vs 0.387 by RWMN) on the validation set and 0.373 (vs 0.363) on the test set. Benefiting from such modeling ability, FVTA consistently outperforms the classical attention models including soft attention, MCB [13] and TGIF [20]. The result demonstrates the consistent advantages of FVTA over other attention models in question-answering for multiple sequence data.

| Method | Val | Test |
|---|---|---|
| SSCB [50] | 0.219 | - |
| MemN2N [50] | 0.342 | - |
| DEMN [27] | - | 0.300 |
| Soft Attention | 0.321 | - |
| MCB [13] | 0.362 | - |
| TGIF Temporal [20] | 0.371 | - |
| RWMN [39] | 0.387 | 0.363 |
| FVTA | **0.410** | **0.373** |

TABLE 8: Accuracy comparison on the test and the validation set of the MovieQA dataset. The test set performance can only be evaluated on the MovieQA server, and thus not all the studies provide the accuracy on Test set.

Fig. 10 illustrates the output of our FVTA model. FVTA can not only predict the correct answer, but also identify the most relevant subtitle description as well as the movie clip frames. As shown in Fig. 10, FVTA can provide fine-grained level justifications such as the most informative movie frames or subtitle sentences, whereas most of the existing methods cannot find fine-grained justifications from the attention computed at the movie clip level. We believe the results show the benefits and potentials of FVTA model.

## 7 Conclusions and Future Work

In this paper, we introduced a new VQA task and presented the first public dataset of real personal photo albums to the research community to study this interesting problem. To address this problem, we presented a novel neural network model called Focal Visual-Text Attention network for answering questions over visual-text sequences. FVTA employed a hierarchical process to dynamically determine which modality and snippets to focus on in the sequential data to answer the question, and hence can not only predict the correct answers but also find the correct supporting justifications to help users verify the system's results. The comprehensive experimental results demonstrated that FVTA achieves comparable or even better than state-of-the-art results on the proposed dataset as well as the MovieQA benchmark of sequential visual-text data. We consider our work as the first step towards solving this new QA problem, establishing the experimental benchmark for future research to explore. Our future work includes improving the scalability of FVTA and extending FVTA to large scale long visual-text sequences.

4. The best test accuracy on the leaderboard by the time of paper submission (Aug. 2018) is 0.42 (A2A: Attention to Attenion). It is not included in the table as there is no publication to cite.

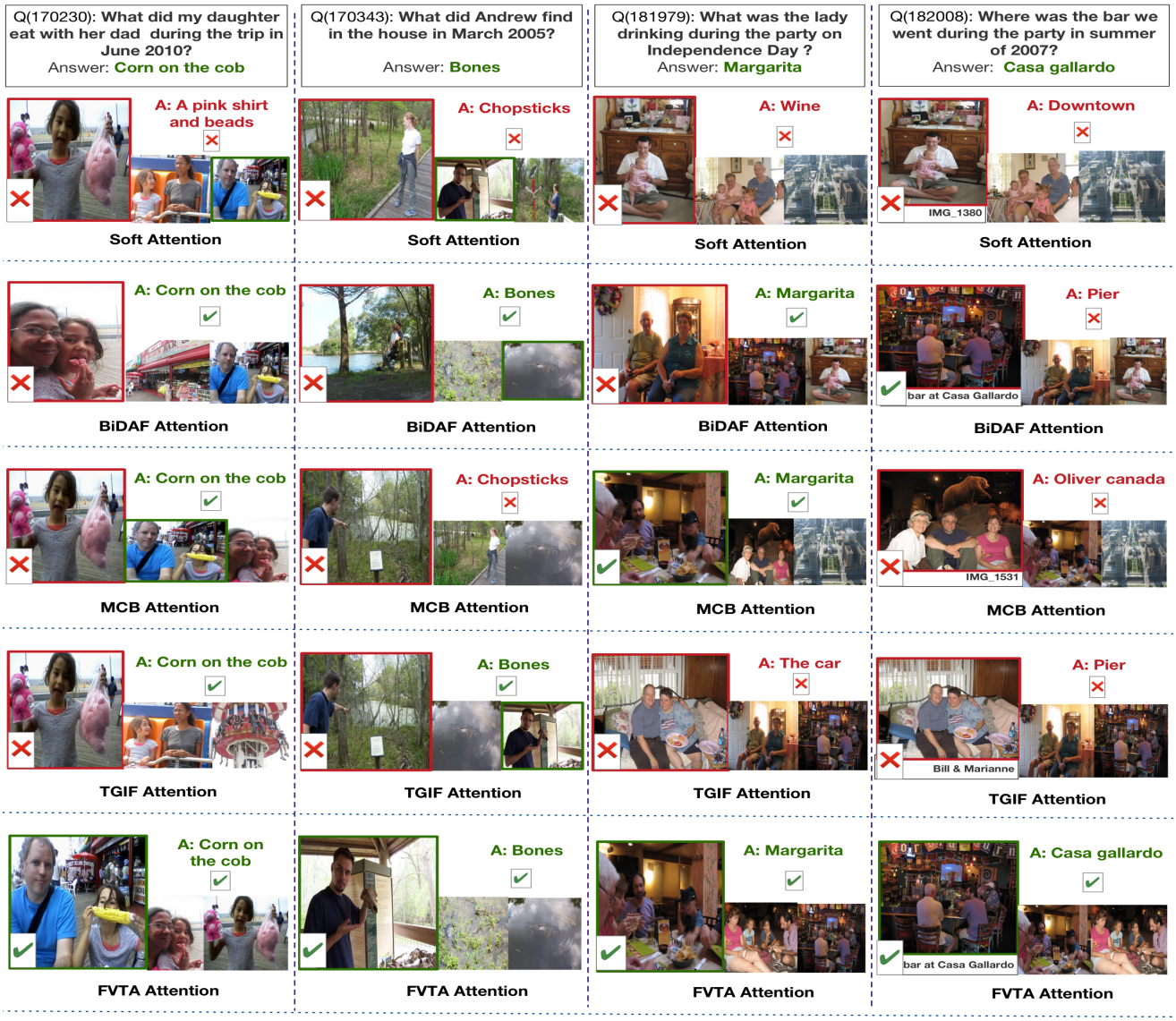



Fig. 9: Qualitative comparison of FVTA model and other attention models on the MemexQA dataset. For each question, we show the answer and the images of the highest attention weights. Images are ranked from left to right based on the attention weights. The correct images and answers have green border whereas the incorrect ones are surrounded by the red border.

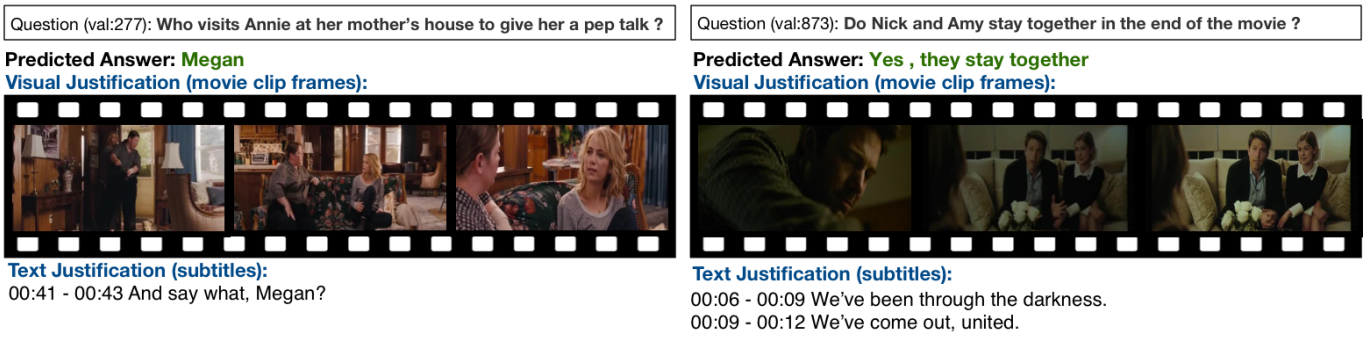



Fig. 10: Qualitative analysis of FVTA on the MovieQA dataset. It shows the visual justification (movie clip frames) and text justification (subtitles) based on the top attention activation. Both justifications provide supporting evidence for the system to get the correct answer.

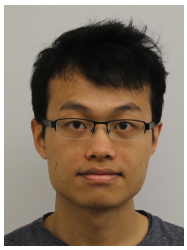
**Acknowledgements** We would like to thank anonymous reviewers as well as our colleagues Yale Song and Sachin Farfade for their useful comments and suggestions. Google Cloud provided GCP research credits for computation. This work was partially supported by Yahoo InMind project and Flickr Computer Vision and Machine Learning group. Yannis' work was performed while

he was at Yahoo Research. This work was partially supported by the financial assistance award 60NANB17D156 from U.S. Department of Commerce, National Institute of Standards and Technology (NIST). This work was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00340. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation/herein. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, NIST, DOI/IBC, or the U.S. Government.

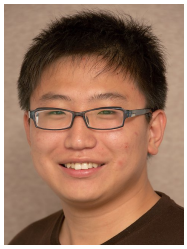
## REFERENCES

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Dont just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.
[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, 2016.
[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.
[5] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. *arXiv preprint arXiv:1705.06676*, 2017.
[6] V. Bush. The atlantic monthly. *As we may think*, 176(1):101–108, 1945.
[7] L. Cao, J. Luo, and T. S. Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *MM*, 2008.
[8] L. Cao, J. Luo, H. A. Kautz, and T. S. Huang. Annotating collections of photos using hierarchical event and scene models. In *CVPR*, 2008.
[9] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *SIGCHI*, 2007.
[10] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 2017.
[11] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied question answering. 2018.
[12] S. Davies. Still building the memex. *Communications of the ACM*, 54(2):80–88, 2011.
[13] 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.
[14] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *NIPS*, 2015.
[15] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi. Iqa: Visual question answering in interactive environments. 2018.
[16] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.
[17] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. 2018.
[18] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015.
[19] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, J. Devlin, A. Agrawal, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *NAACL*, 2016.
[20] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.
[21] L. Jiang, L. Cao, Y. Kalantidis, S. Farfade, J. Tang, and A. G. Hauptmann. Delving deep into personal photo and video search. In *WSDM*, 2017.
[22] L. Jiang, J. Liang, L. Cao, Y. Kalantidis, S. Farfade, and A. G. Hauptmann. Memexqa: Visual memex question answering. *arXiv:1708.01336*, 2017.
[23] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ICMR*, 2015.
[24] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.
[25] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017.
[26] K. Kafle, B. Price, S. Cohen, and C. Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.
[27] K.-M. Kim, M.-O. Heo, S.-H. Choi, and B.-T. Zhang. Deepstory: video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*, 2017.
[28] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
[29] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
[30] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua. A multi-level contextual model for person recognition in photo albums. In *CVPR*, 2016.
[31] Y. Li, G. Lin, B. Zhuang, L. Liu, C. Shen, and A. van den Hengel. Sequential person recognition in photo albums with a recurrent network. In *CVPR*, 2017.
[32] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. Hauptmann. Focal visual-text attention for visual question answering. In *CVPR*, 2018.
[33] D. Lin, A. Kapoor, G. Hua, and S. Baker. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *ECCV*, 2010.
[34] D. Liu, M. Wang, X.-S. Hua, and H.-J. Zhang. Smart batch tagging of photo albums. In *MM*, 2009.
[35] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun. ivqa: Inverse visual question answering. 2018.
[36] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.
[37] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.
[38] A. Miech, I. Laptev, and J. Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.
[39] S. Na, S. Lee, J. Kim, and G. Kim. A read-write memory network for movie story understanding. *arXiv preprint arXiv:1709.09345*, 2017.
[40] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016.
[41] D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. *arXiv preprint arXiv:1804.00775*, 2018.
[42] B. Patro and V. P. Namboodiri. Differential attention for visual question answering. In *CVPR*, 2018.
[43] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.
[44] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
[45] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015.
[46] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.
[47] Y. Shen, P.-S. Huang, J. Gao, and W. Chen. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055. ACM, 2017.
[48] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.
[49] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.
[50] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016.
[51] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016.
[52] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. R. Dick. FVQA: fact-based visual question answering. *CoRR*, abs/1606.05433, 2016.
[53] S. Wang and J. Jiang. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747*, 2016.
[54] S. Wang and J. Jiang. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*, 2016.
[55] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. 2016.
[56] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video

question answering via gradually refined attention over appearance and motion. In *MM*, 2017.
[57] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.
[58] H. Yang, L. Chaisorn, Y. Zhao, S.-Y. Neo, and T.-S. Chua. Videoqa: question answering on news video. In *MM*, 2003.
[59] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.
[60] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, 2015.
[61] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
[62] M. Zhao, Y. W. Teo, S. Liu, T.-S. Chua, and R. Jain. Automatic person annotation of family photo album. In *CIVR*, 2006.
[63] C. Zhu, F. Wen, and J. Sun. A rank-order distance based clustering algorithm for face tagging. In *CVPR*, 2011.
[64] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann. Uncovering temporal context for video question and answering. *arXiv preprint arXiv:1511.04670*, 2015.
[65] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. 0swering in images. In *CVPR*, 2016.

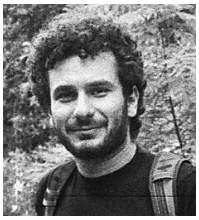


**Junwei Liang** is a Ph.D. candidate in the Language Technologies Institute at Carnegie Mellon University. received the graduation degree from Renmin University, Beijing, China, in 2015, where he worked as a research assistant in the multimedia lab at Renmin University. He received the Master of Language Technologies degree from the Language Technologies Institute at Carnegie Mellon University in 2017. He was at Google Inc., Pittsburgh, PA, as a student researcher in 2018. His research interests include computer vision, natural language processing and machine learning.



**Lu Jiang** received his Ph.D degree in Computer Science from Carnegie Mellon University (CMU) in 2017, and the M.Sc. and B.Eng. from Xi'an Jiaotong University. Currently, he is a research scientist at Google AI. His research is focused on multimedia, video understanding, vision and language, and machine learning.

**Liangliang Cao** is a Research Associate Professor at UMass Amherst. He is also a cofounder of HelloVera AI. He taught at Columbia University an adjunct professor from 2013 to 2018. Previously he worked as a senior scientist at Yahoo Labs and earlier a research staff member at IBM Watson Research Center. He won the 1st place of ImageNet LSVRC Challenge in 2010. He is an IEEE senior member and a recipient of ACM SIGMM Rising Star Award. His research interests include computer vision, speech, and language.



**Yannis Kalantidis** is a research scientist at Facebook Research in Menlo Park, California. He grew up in Athens, Greece and lived there till 2015, with brief breaks in Sweden, Spain and the United States. He got his PhD on large-scale search and clustering from the National Technical University of Athens in 2014. He was a postdoc and research scientist at Yahoo Research in San Francisco for two years, leading the visual similarity search project at Flickr and participated in the Visual Genome dataset project with Stanford. He is currently conducting research on video understanding, representation learning and modeling of vision and language.

**Jia Li** holds the position of AI Fellow and Adjunct Professor at Stanford University in the School of Medicine. In Healthcare, she is interested in how AI could improve the outcomes of individual patients as well as hospitals. She was the Head of R&D at Google Cloud AI. Their mission is to democratize AI and advance AI. Her org focus on both research innovation to solve real world problems and developing the full stack of AI products on Google Cloud to power solutions for diverse industries. Before joining Google, She was the Head of Research at Snap, leading the research innovation effort. Before Snap, she led the Visual Computing and Learning Group at Yahoo! Labs. In 2014, she was selected to receive the Super Star award at Yahoo!, the highest award at the company. she was also awarded the Master Inventor Award for her innovations in AI/ML. She received her Ph.D. degree from the Computer Science Department at Stanford University. She was the leader of the OPTIMOL team, which won the first prize in the Semantic Robotics Vision Challenge sponsored by NSF and AAAI in 2007. She served as the Program Chair of ACM Multimedia 2017, Area Chair of ICCV 2017 and CVPR 2019, Industry Relationship Chair of CVPR 2016 and Volunteers Chair of CVPR 2010. She is on The Computer Vision Foundation Industrial Advisory Board. She is serving as Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) and Associate Editor of the Visual Computer: International Journal of Computer Graphics by Springer. She is selected as a Young Global Leader by the World Economic Forum in 2018. Her work has been reported in the media including: Forbes, TechCrunch, CNBC, New Scientist, MIT Technology Review and more in recent years.

**Alexander G. Hauptmann** received the BA and MA degrees in psychology from Johns Hopkins University, Baltimore, Maryland, the degree in computer science from the Technische Universitat Berlin, Berlin, Germany, in 1984, and the PhD degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, Pennsylvania, in 1991. He is currently with the faculty of the Department of Computer Science and the Language Technologies Institute, CMU. His research interests include several different areas: man-machine communication, natural language processing, speech understanding and synthesis, video analysis, and machine learning. From 1984 to 1994, he was working on speech and machine translation, when he joined the Informedia project for digital video analysis and retrieval, and led the development and evaluation of news-ondemand applications.