Visual Memory QA: Your Personal Photo and Video Search Agent

Lu Jiang[1], **LiangLiang Cao**[2], **Yannis Kalantidis**[2], **Sachin Farfade**[2], **Alexander G. Hauptmann**[1]
[1] Carnegie Mellon University, United States
[2] Yahoo Research, United States
{lujiang, alex}@cs.cmu.edu, liangliang.cao@gmail.com,
ykalant@image.ntua.gr, fsachin@yahoo-inc.com

Abstract

The boom of mobile devices and cloud services has led to an explosion of personal photo and video data. However, due to the missing user-generated metadata such as titles or descriptions, it usually takes a user a lot of swipes to find some video on the cell phone. To solve the problem, we present an innovative idea called Visual Memory QA which allow a user not only to search but also to ask questions about her daily life captured in the personal videos. The proposed system automatically analyzes the content of personal videos without user-generated metadata, and offers a conversational interface to accept and answer questions. To the best of our knowledge, it is the first to answer personal questions discovered in personal photos or videos. The example questions are "what was the lat time we went hiking in the forest near San Francisco?"; "did we have pizza last week?"; "with whom did I have dinner in AAAI 2015?".

Introduction

The prevailing of mobile devices and cloud services has led to an unprecedented growth of personal photo and video data. A recent study shows that the queries over personal photos or videos are usually task- or question-driven (Jiang et al. 2017). For question-driven queries, users seem to be using photos or videos as a mean to recover pieces from their own memories, i.e. looking for a specific name, place or date. For example, a user might ask "what was the last time we went hiking?"; "did we have pizza last week?" or "with whom did I have dinner in AAAI 2015?".

We define the problem of seeking answers about the user's daily life discovered in his or her personal photo and video collection as VMQA (Visual Memory Question Answering). As about 80% of personal photos and videos do not have metadata such as tags or titles (Jiang et al. 2017), this functionality can be very useful in helping users find information in their personal photos and videos. Visual Memory QA is a novel problem and has two key differences from VQA (Visual QA) (Antol et al. 2015): first the user is able to ask questions over a collection of photos or videos in Visual Memory QA as opposed to a single image in VQA. As shown in Fig. 1, given an image it is trivial for an adult to answer a question in VQA. However, it is considerably more difficult for the same adult to answer questions in VMQA. This is particularly difficult to answer questions over a collection videos. Second, the question space in VMQA is a subset of that in VQA, which only includes the questions a user might ask later to recall his or her memories. Because of the two differences, Visual Memory QA is expected to be more useful in practice.

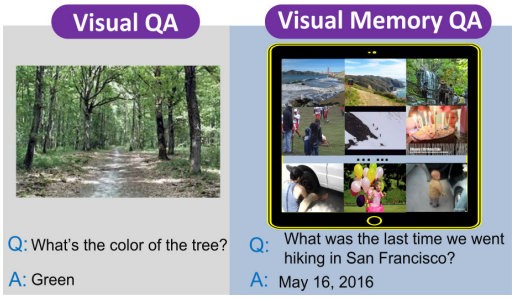

Figure 1: Comparison of Visual QA & Visual Memory QA.

To address this novel problem, this paper introduces a prototype system that can automatically analyzes the content of personal videos/photos without user-generated metadata, and offers a conversational interface to answer questions discovered from the user's personal videos/photos. Technically, it can be regarded as an end-to-end neural network, consisting of three major components: a recurrent neural network to understand the user question, a content-based video engine to analyze and find relevant videos, and a multi-channel attention neural network to extract the answer. To the best of our knowledge, the proposed system is the first to answer personal questions discovered in personal photos or videos.

Visual Memory Question Answering

As shown in Fig. 2, the proposed model is inspired by the classical text QA model (Ferrucci et al. 2010), consisting of three major components: a recurrent neural network to understand the user question, a content-based video engine to find the relevant videos, and a multi-channel attention feed-forward neural network to extract the answer. Each component is pre-trained on its own task, and then the first and the third components are fine-tuned on our annotated benchmark data by Back Propagation.

In the recurrent neural network, the task is to understand

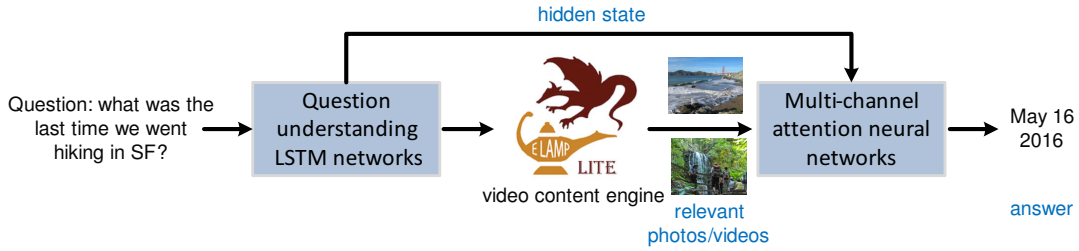

Figure 2: Framework of the proposed Visual Memory QA system.

the question and classify it into a predefined answer type. We predefine a set of question and answer types based on their frequencies in Flickr visual search logs (Jiang et al. 2017). See Table 1. A two-layer LSTM neural network is incorporated as the classifier where the embedding of each word in the question is sequentially fed into the LSTM units. As the answer types are mutually exclusive, a softmax logistic loss is employed to train the network. Besides, this question understanding component is also responsible for parsing the question to extract the named entity (person, organization, place and time).

Table 1: Question and answer types in the proposed system.

Question Type	Answer Type	Example
which	photo, video	show me the photo of my dog?
when	date, year, season, hour, etc.	What was the last time we went hiking?
where	scene, gps, city, country, etc.	Where was my brother's graduation ceremony in 2013?
what	action, object, activity, etc.	What did we play during this spring break?
who	name, face, etc.	With whom did I have dinner in AAAI 2015?
how many	number	How many times have I had sushi last month?
yes/no	yes, no	Did I do yoga yesterday?

The second component is a content video/photo engine that can automatically understand and index personal videos purely based on the video content. It takes a natural language sentence as the input, and outputs a list of semantically relevant videos, i.e. text-to-video (Jiang et al. 2015a). The top ranked relevant videos are fed into the third component. We employ a state-of-the-art engine called E-Lamp Lite which not only provides accurate video search and understanding but also can scale up to 100 million videos (Jiang et al. 2015b).

The last component is a neural network to extract the answer. It receives, from the question understanding network, a hidden state that embeds the information about the predicted answer type, and the top ranked relevant videos from the video content engine. Each relevant video is associated with information organized into channels, such as the timestamp, the action concepts, scene concepts, object concepts and, in some cases, the GPS coordinates. The task now switches to localizing the answer in the multiple input channels. For example, the attention should be on timestamp for "when" questions, and on food concepts for "what did we eat" questions. This is now achieved by a multi-channel attention feed-forward neural network. For the current prototype, a few manual templates are also employed to further improve the accuracy.

Demonstration

The demonstration will be organized in two phases: *a)* a brief introduction, and *b)* a hands-on phase. In *a)*, the main features of the Visual Memory QA system will be explained and some example queries will be demonstrated. In *b)*, the public is invited to interact directly with the system and test its capabilities over a laptop or on a cell phone. Specifically. all of the personal videos in YFCC dataset (about 0.8M) will be employed as a giant collection of a single anonymous user, and the public can ask questions and examine results in less than 2 seconds. The demo will be running on a laptop and we will bring cell phones and other laptops to show the results. No additional devices are needed for this demo.

Conclusions and Future Work

This demo paper presents a novel and promising Visual Memory QA system, an intelligent agent or chatbot that can answer questions about users' daily lives discovered in their personal photos and videos. We have developed a prototype system that can efficiently answer questions over 1 million personal videos. We are still working on obtaining more annotated data to qualitatively evaluate the accuracy of on the end-to-end task. In the future, we plan to release a benchmark on this novel and interesting problem.

Acknowledgments

This work was partially supported by Yahoo InMind Project and the IARPA via Department of Interior National Business Center contract number D11PC20068.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; et al. 2010. Building watson: An overview of the deepqa project. *AI magazine* 31(3):59–79.

Jiang, L.; Yu, S.-I.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2015a. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ICMR*.

Jiang, L.; Yu, S.-I.; Meng, D.; Yang, Y.; Mitamura, T.; and Hauptmann, A. G. 2015b. Fast and accurate content-based semantic search in 100m internet videos. In *MM*.

Jiang, L.; Cao, L.; Kalantidis, Y.; Farfade, S.; Tang, J.; and Hauptmann, A. G. 2017. Delving deep into personal photo and video search. In *WSDM*.