Journal of Visual Communication and Image Representation 55 (2018) 40–55

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Deepdiary: Lifelogging image captioning and summarization☆

Chenyou Fan*, Zehua Zhang, David J. Crandall

School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

ARTICLE INFO

Keywords:
Lifelogging
First-person
Image captioning
Diary
Privacy

ABSTRACT

Automatic image captioning has been studied extensively over the last few years, driven by breakthroughs in deep learning-based image-to-text translation models. However, most of this work has considered captioning web images from standard data sets like MS-COCO, and has considered single images in isolation. To what extent can automatic captioning models learn finer-grained contextual information specific to a given person's day-to-day visual experiences? In this paper, we consider captioning image sequences collected from wearable, life-logging cameras. Automatically-generated captions could help people find and recall photos among their large-scale life-logging photo collections, or even to produce textual "diaries" that summarize their day. But unlike web images, photos from wearable cameras are often blurry and poorly composed, without an obvious single subject. Their content also tends to be highly dependent on the context and characteristics of the particular camera wearer. To address these challenges, we introduce a technique to jointly caption sequences of photos, which allows captions to take advantage of temporal constraints and evidence across time, and we introduce a technique to increase the diversity of generated captions, so that they can describe a photo from multiple perspectives (e.g., first-person versus third-person). To test these techniques, we collect a dataset of about 8000 realistic lifelogging images, a subset of which are annotated with nearly 5000 human-generated reference sentences. We evaluate the quality of image captions both quantitatively and qualitatively using Amazon Mechanical Turk, finding that while these algorithms are not perfect, they could be an important step towards helping to organize and summarize lifelogging photos.

1. Introduction

Wearable devices like FitBit and Apple Watch have become very popular [41], allowing people to collect fine-grained information about their daily lives, such as step counts, heart rates, calories burned, etc. While these devices have proven to be powerful motivators for weight loss and other personal goals [13], they are limited in the types of activity that they can record. Capturing the complex dynamics of one's day, such as the people and objects that one interacts with, requires higher fidelity sensor information.

Wearable *lifelogging cameras* offer the promise of recording this richer information, by taking photos at regular intervals to create a visual, first-person record of one's visual experiences throughout the day [21,36]. These cameras have long been studied in the ubiquitous computing research community for applications like recording food and other health choices, treating dementia, assisting people with visual impairments, and quantifying social interactions [1,4,22,38,44,63,68]. However, beyond a few niche applications such as documenting police officers' interactions with the public [66], lifelogging cameras have failed to catch on. For example, Narrative Clip and Autographer (Fig. 1) were lifelogging devices that generally received strong positive reviews for their technology [25,82], but went out of business within a few years [9] when consumer demand failed to materialize.

In practice, a major problem with these lifelogging cameras is that while the rich data they collect could enable interesting applications, the data can also overwhelm the average user [9]. The Narrative Clip, for example, took photos at 30 s intervals and easily generated over a thousand photos per day. These photo streams are typically highly repetitious, with hundreds of photos of mundane moments, many of which are blurry, poorly composed, or just boring. Manually reviewing these photos on a daily basis to find photos worth sharing with friends or on which to perform interesting analysis can be intractable for all but the most enthusiastic of users. Meanwhile, many first-person photos contain private or potentially embarrassing content, including shots of sensitive records and documents (e.g., ATM card numbers, private emails or text messages on a smartphone display, etc.) and private personal activities or interactions (e.g., in bathrooms, bedrooms, and locker rooms) [39]. Newer wearable cameras such as Snapchat

☆ This paper has been recommended for acceptance by Petia Radeva.

* Corresponding author.
E-mail addresses: fan6@indiana.edu (C. Fan), zehzhang@indiana.edu (Z. Zhang), djcran@indiana.edu (D.J. Crandall).

https://doi.org/10.1016/j.jvcir.2018.05.008
Received 14 March 2017; Received in revised form 7 May 2018; Accepted 10 May 2018
Available online 23 May 2018

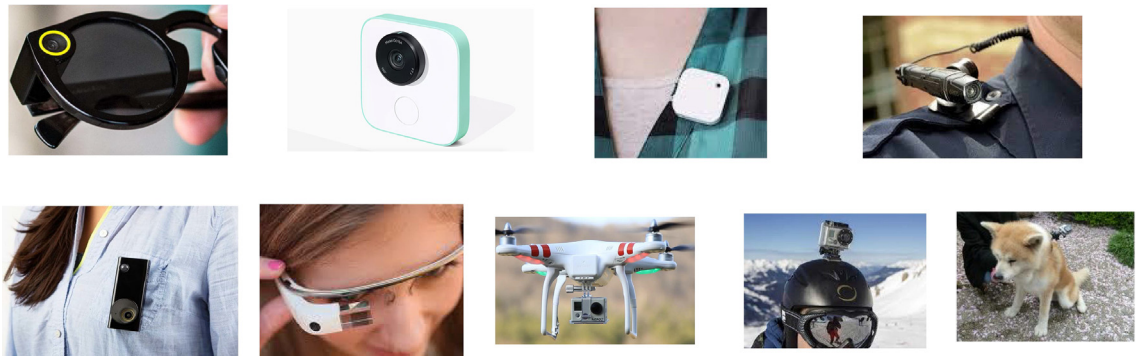


Fig. 1. *Examples of wearable cameras:* Top row: Snap Spectacles with integrated video camera, Google Clips, Narrative Clip lifelogging camera, police body camera; Bottom row: Autographer lifelogging camera, Google Glass, and GoPro Hero cameras mounted on drones, people, and pets.

Spectacles and Google Clips try to address these problems by taking a photo only when requested by the user (as with the Spectacles) or according to an algorithm that tries to detect interesting scene content (as with Clips). But these approaches compromise the major promise of life-logging cameras: collecting unbiased, dense samples of one's day.

To make lifelogging with wearable cameras tractable for the mainstream consumer, we need new automatic techniques for first-person photo organization, including helping people find both the good photos they might like to share and the private photos they may want to delete, as well as helping to summarize their huge photo collections in more manageable ways. Some very recent work has begun to address the challenges of automatically organizing first-person imagery from several different perspectives [10,11]. Work in computer vision has applied traditional supervised recognition techniques to this problem in order to detect pre-defined sets of objects [30,52], scene types [31], and places [86]. Other work has studied recognizing what is happening in first-person imagery, including recognizing pre-defined activities and actions of the camera wearer [5,14,29,76] or of others in the scene [77]. However, a problem with these techniques is that the space of possible objects and activities is so huge that it may be very difficult to enumerate and run enough classifiers to extract all meaningful content. An alternative is to use unsupervised techniques that do not try to recognize specific content, but instead look for recurring visual patterns that may correspond to important content [57,81], or choose key images to summarize video or photo streams [62,73,90]. These techniques can identify repeated patterns in imagery without explicit training, but cannot produce the semantic-level information that classifiers can.

What if visual lifelogs could be distilled into automatic textual summaries, or "narrations," of a person's day? Textual representations could be a natural, flexible way to summarize large-scale visual lifelogs for the average user, while enabling interesting applications such as automatically generating textual diaries to convey the "story" of a lifelogging image stream. While automatic image captioning has become a well-studied research area in recent years [46,48], the vast majority of this work is focused on datasets of images from the web, like MS-COCO [59] and ImageNet [53]. Lifelogging photos streams have several major differences as compared to web images, however. Because they are taken automatically and indiscriminately, lifelogging photos often suffer from blur or poor illumination, and unlike web images or traditional consumer photographs, they usually do not focus on a prominent subject that can neatly be described by a single caption. Moreover, a lifelogging photo is not taken in isolation, but is instead part of a stream of related photos. Thus it does not make sense to caption individual images, but instead to caption sets of photos, using contextual information across different photos to produce more informative and useful captions.

In this paper, we develop and evaluate techniques for producing automatic textual summaries of lifelogging photo streams. To do this, we adapt existing deep captioning techniques [46] to the lifelogging domain in several ways. First, to deal with images that are noisier and lack prominent subjects, we propose a new strategy to try to encourage diversity in the sentences, which we found to be particularly useful in describing lifelogging images from different perspectives. Second, instead of simply captioning individual images, we consider the novel problem of jointly captioning lifelogging streams, i.e. generating captions for temporally-contiguous groups of photos corresponding to coherent activities or scene types, and then "summarizing" the group of photos with a single best caption. Not only does this produce a more compact and potentially useful organization of a user's photo collection, but it also could create an automatically-generated textual summary or "diary" of a user's day based only on their photos. The sentences themselves are also useful to aid in image retrieval by keyword search, which we illustrate for the specific application of searching for potentially private images (e.g., containing keywords like "bathroom"). Joint caption estimation over multiple images also reduces noise and errors in the captioning results, since evidence from multiple photos is used to infer each caption. We formulate this joint captioning problem in a Markov Random Field model and show how to solve it efficiently.

To train and test these techniques, we created a lifelogging photo collection and labeling website to collect realistic photos and annotations from lifelogging users. Several thousand photos from two of the authors' lifelogs were collected, and over 800 were manually annotated with more than 4,000 captions by both the authors and Mechanical Turk users for training and testing purposes. We evaluate the performance of image captioning both qualitatively and quantitatively on this set.

In summary, our contributions are fivefold: (1) we learn and apply deep image captioning models to lifelogging photos; (2) we propose a novel method for generating photo descriptions with diverse structures and perspectives; (3) we propose a novel technique for inferring captions for streams of photos taken over time in order to find and summarize coherent activities and other groups of photos; (4) we create an online framework for collecting and annotating lifelogging images, and use it to collect a realistic lifelogging dataset consisting of thousands of photos and thousands of reference sentences, which we have publicly released;[1] and (5) we evaluate our techniques on our data, both quantitatively and qualitatively, under different simulated use cases. A preliminary version of this work appeared in [17], which first proposed the idea of captioning in lifelogging images.

2. Related work

For many years, wearable devices have been studied as a means of collecting various types of data (GPS, accelerometer, heart rate, etc.) about people as they go about their daily lives. Applications have

[1] http://vision.soic.indiana.edu/deepdiary.

included helping users manage daily tasks [85], monitor financial choices [84], improve health behavior [49,68,70], etc. As a special case, wearable *cameras* have been explored for over a decade in the research community [4,38,63], but only recently have become practical enough for consumers to use on a daily basis. The rich data that cameras can collect has inspired ideas for a variety of new applications, including aiding human memory for retrospection [16,21,36,93], helping students learn [7], assisting people with visual impairments [45], studying human behavior for psychological studies [44,22,5], and so on.

These wearable camera applications raise a number of challenges. From a privacy and security perspective, for example, Denning et al. [20] and Nguyen et al. [69] study how bystanders react to wearable cameras, while Hoyle et al. [39] identify privacy risks to the camera wearers themselves. These studies raise a variety of concerns about how wearable cameras may alter societal interactions and our perceptions of privacy, and raise legal questions about reasonable expectations of privacy in a camera-rich world; in contrast, others view wearable cameras as a potentially powerful force for "democratization" [63], since wearable camera data is recorded and held by individual citizens, as opposed to traditional surveillance video which is typically controlled by governments and institutions. From a technical standpoint, many applications would require automatic techniques to analyze and organize the vast quantities of images that wearable cameras collect. In the computer vision field, recent work has begun to study this new style of imagery, which is significantly different from photos taken by traditional point-and-shoot cameras. Specific research topics have included recognizing objects [30,52], scenes [31], and activities [29,14,76,77]. Some computer vision work has specifically tried to address privacy concerns, by recognizing photos taken in potentially sensitive places like bathrooms [86], or containing sensitive objects like computer monitors [52]. However, these techniques typically require that classifiers be explicitly trained for each object, scene type, or activity of interest, which limits their scalability.

Instead of classifying lifelogging images into pre-defined and discrete categories, we propose to annotate them with automatically-generated, free-form image captions, inspired by recent progress in deep learning. Convolutional Neural Networks (CNNs) have recently emerged as powerful models for object recognition in computer vision [26,32,53,83], while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) have been developed for learning models of sequential data, like natural language sentences [24,34]. While these networks were initially applied to sequence-to-sequence mapping problems like translating from one language to another, the combination of CNNs for recognizing image content and RNNs for modeling language has been shown to generate surprisingly rich image descriptions [46,65,89], essentially "translating" from image features to English sentences [47]. Besides employing neural network models for caption generation, Kulkarni et al. [54] jointly reason about objects and attributes through a CRF and fill corresponding slots in a template. Farhadi et al. [28] retrieve the caption from a database based on the templated semantic representation computed as a triplet. A syntactically well-formed tree is utilized in [55,67] as a more effective linguistic template. However, the captions generated by these methods are either not natural sounding enough or not able to comprehensively describe the image.

Deep image captioning has thus become an active research area over the last two years [15,23,46,50,64,89]. For instance, several lines of recent work try to achieve finer-grained analysis that identify the specific image regions that are "responsible" for producing certain words [27,61,75,91,92,94]. Fang et al. [27], for example, generate words that are likely to appear in an image caption by using a "word detector" – a CNN that represents images as feature maps and produces word probabilities by integrating information from multiple image regions. Then they formulate sentence generation as an optimization problem which searches for sentences of high likelihood conditioned on detected words. Xu et al. [91] propose a model based on visual attention which implicitly learns latent alignment of visual concepts and words without explicitly detecting objects. They use an LSTM to produce sentences by generating one word at a time, in which the next predicted word is conditioned on a context vector computed as a weighted sum of image region vectors. The weights themselves are learned and interpreted as "attention," or the probability that each location is where the network should look for generating the next word. Yang et al. [92] extend the standard encoder-decoder framework for image captioning by adding a number of "review" steps with an attention mechanism similar to Xu et al. [91], except that they produce all attentive image representations before the decoding stage, using the intuition that the order of attention on parts of an image is not constrained by the order of words in a corresponding sentence. Park et al. [72] propose to generate personalized captions by taking prior knowledge about a particular user into account, such as her or his typical vocabulary. They achieved this by adding a memory module to store such prior knowledge.

More recently, image captioning techniques have been applied to visual question answering (VQA) [2,3,19,33,40]. Similar to image captioning, VQA usually uses an image encoder to understand the visual input as well as an answering module to produce textual output, but with an additional question encoder which converts a sentence to fixed length features. Some work [3,40] also employs neural module networks to further decompose the sentence into different linguistic substructures.

Lifelogging images tend to be noisier and more poorly composed than consumer-style photos. We propose two strategies to help counter this challenge. First, we use the fact that lifelogging images are typically not captured in isolation but instead captured in a sequence, e.g. one photo every 30 s for the Narrative Clip camera. This means that simply creating a caption independently for each photo would create an unnecessarily large and redundant set of sentences, but it also means that evidence from multiple photos can be used to create better captions. We thus consider the problem of "summarizing" a sequence of images by assigning captions to automatically-selected groups of photos, such that the number of captions produced is much smaller than the number of photos, but the quality of the captions is significantly greater. Related to our approach is the problem of generating textual descriptions from videos, which can be viewed as sequences of images. Venugopalan et al. [88] use an image captioning model to generate video descriptions from a sequence of video frames. Like previous image captioning papers, their method estimates a single sentence for each sequence, while we explicitly generate multiple diverse sentences and evaluate the image-sentence matching quality to improve the captions from noisy, poorly-composed lifelogging images. Zhu et al. [95] use neural sentence embedding to model a sentence-sentence similarity function, and use LSTMs to model image-sentence similarity in order to align subtitles of movies with sentences from the original books. Their main purpose is to find corresponding movie clips and book paragraphs based on visual and semantic patterns, whereas ours is to infer novel sentences from new lifelogging image streams.

Our second strategy for addressing the challenges of lifelogging photos is to explicitly encourage diversity in our generated captions. This problem can be interpreted as finding a set of highly probable, yet distinct, sentences. Most captioning approaches simply return the top few highest-likelihood sentences, which we found to be very similar. We use diverse *m*-best solutions, a technique [8] originally intended to find diverse solutions in probabilistic models such as Markov Random Field (MRF). The idea is to perform inference multiple times, each time finding the optimal solution that is sufficiently different from all previously found solutions. There is also recent work [56] aimed at training deep networks to produce diverse outputs by using an ensemble set of predictors which capture different data distributions. Their work also showed an ability to generate different sentences. The difference between our work and theirs is that ours has a probabilistic

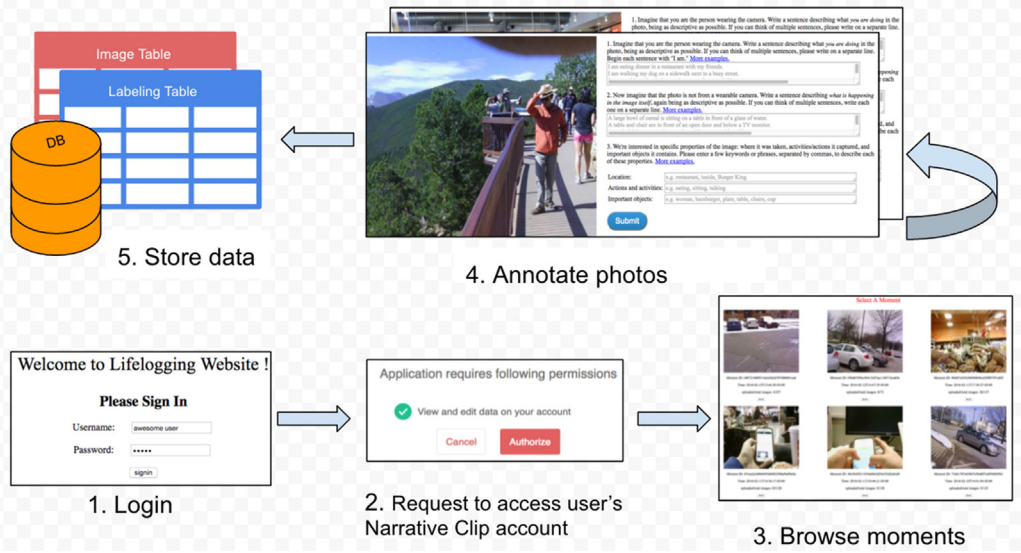


Fig. 2. Flow chart of out lifelogging image and reference sentence collection procedure.

interpretation and can manipulate the word prior in order to generate sentences of different prefixes.

Perhaps most related to our approach is the very recent work of Bolaños et al. [12], who also caption egocentric imagery. They also use temporal constraints, but focus on long-range temporal connections between events, whereas we take the complimentary approach of using shorter-term constraints within individual events. We also generate multiple sentences per image using a technique that explicitly encourages diversity, to allow it to describe images from multiple perspectives (e.g. first- vs third-person).

3. Lifelogging data collection

We collected a new dataset because there were no existing lifelogging or egocentric image datasets with caption annotations (although we are aware of one that is forthcoming [12]). Two of the authors wore Narrative Clip lifelogging cameras over a period of about five months (June-Aug 2015 and Jan-Feb 2016), to create a repository of 7,716 lifelogging photos. To facilitate collecting lifelogging photos and annotations, we built a website which allowed users to upload and label photos in a unified framework, using the Narrative Clip API.[2] Fig. 2 shows an overview of the data collection procedure. At the end of each day, users connected the Narrative Clip to their computer, and Narrative's software processed the day's photos and uploaded them to the user's private Narrative account. They then accessed our online system, which requested authorization to log into their Narrative account via the API. The system showed users batches of photos (which Narrative calls "Moments") that the API identified as being related to one another (using metadata like light level, timestamp, and GPS). Users selected the moment(s) they wanted to include in our dataset; this opt-in mechanism allowed them to exclude individual photos and/or whole moments that they did not want to be included in the dataset (e.g., due to privacy concerns).

We collected textual annotations for training and testing the system in two different ways. First, the two authors and four of their friends and family members used the online system to submit sentences for randomly-selected images, producing 2,683 sentences for 696 images. Annotators were asked to produce at least two sentences per image: one that described the photo from a first-person perspective (e.g., "I am eating cereal at the kitchen table.") and one from a third-person perspective (e.g., "A bowl of cereal sits on a kitchen table."). We requested sentences from each of these perspectives because we have observed that some scenes are more naturally described by one perspective or the other. Annotators were welcome to enter multiple sentences, and each image was viewed by an average of 1.45 labelers.

Second, to generate more diversity in annotators and annotations, we published 293 random images[3] on Amazon's Mechanical Turk (AMT), showing each photo to at least five annotators and, as before, asking each annotator to give at least one first-person and one third-person sentence. This produced a set of 4,299 sentences, or an average of 14.7 sentences per image. A total of 121 distinct Mechanical Turk users contributed sentences (ranging from 1 to 97 images per Turk users, with a median of 5). To encourage high-quality responses, we restricted the workers to those living in the United States (to help ensure that they were proficient English speakers) with at least 100 completed Mechanical Turk HITs and a historical approval rating of at least 97%. In a pilot study, we found that it takes between about 30–45 s to write two sentences for an image, so we set the Mechanical Turk pay at 10 cents per image (corresponding to an hourly rate of about US $8–$12). We manually reviewed the responses but because our goal was simply to create as large and diverse a set of captions as possible, we performed minimal filtering, removing just 16 of the 1481 responses that were blank or otherwise clearly not written in good faith. Fig. 3 shows our online labeling interface, while Fig. 4 shows a few random images and sample human annotations from our dataset.

Finally, we also downloaded COCO [59], a popular publicly-available dataset of 80,000 photos and 400,000 sentences. These images are from the Internet and social media sources, and thus are significantly different from the lifelogging context we consider here, but we hypothesized that these images may nonetheless be useful in augmenting our smaller lifelogging training dataset.

4. Automatic lifelogging image captioning

We now present our technique for using deep learning to automatically annotate lifelogging images with captions. We first give a brief review of deep image captioning models, and then show how to take advantage of *streams* of lifelogging images by estimating captions jointly across time, which not only helps reduce noise in captions by enforcing temporal consistency, but also helps summarize large photo

[2] https://open-staging.getnarrative.com/api-docs.

[3] We randomly chose 300, but removed 7 that we were not comfortable sharing with the public (e.g., photos of strangers whose permission we were not able to obtain).

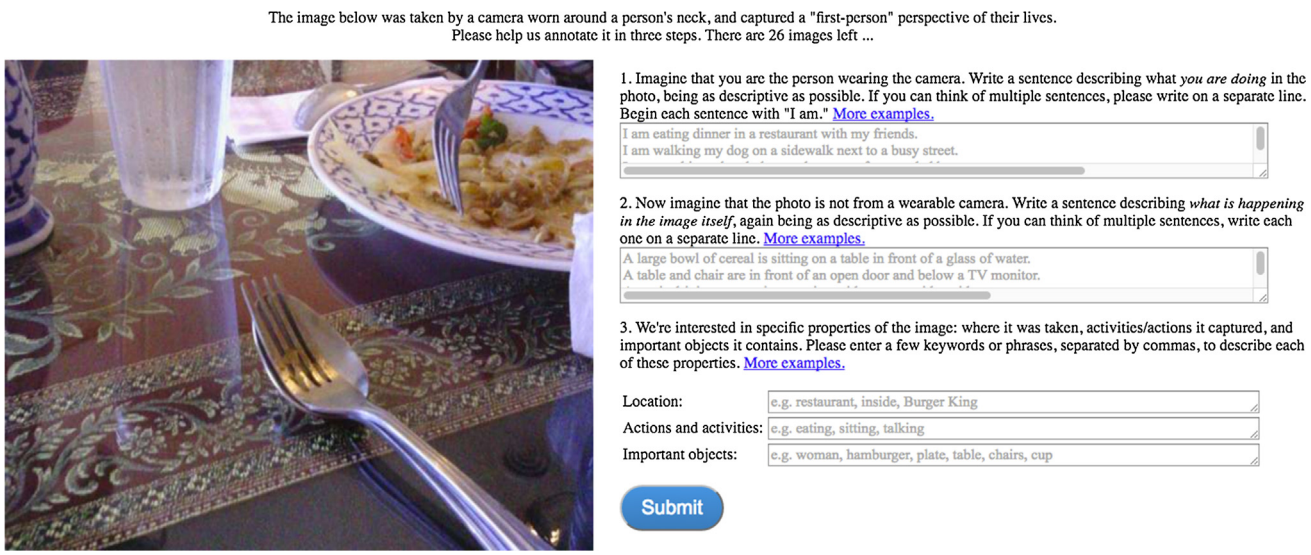


Fig. 3. Online interface for collecting reference sentences from human annotators.

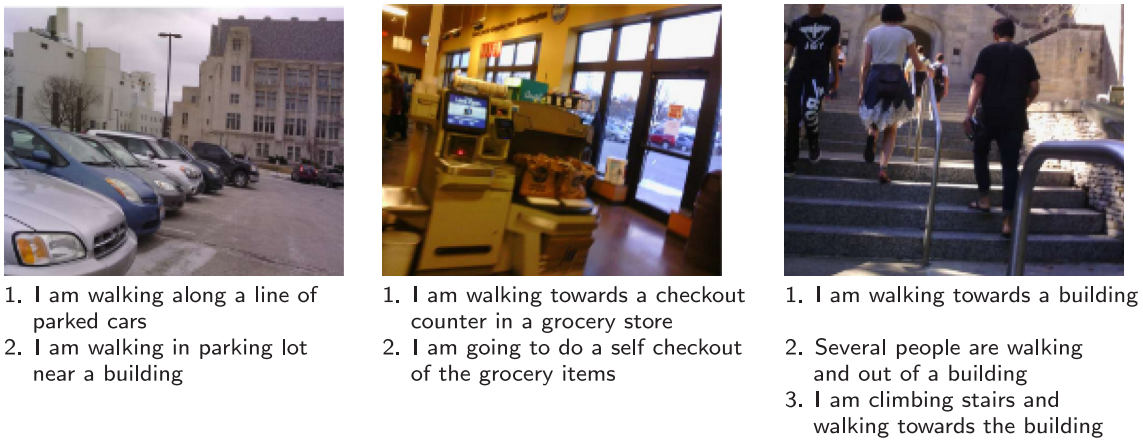


Fig. 4. Sample images and reference sentences from our dataset.

collections with smaller subsets of sentences.

4.1. Background: Deep networks for captioning

Automatic image captioning is a difficult task because it requires not only identifying important objects and actions, but also describing them in natural language. However, recent work in deep learning has demonstrated impressive results in generating image and video descriptions [46,88,95]. The basic high-level idea is to learn a common feature space that is shared by both images and words. Then, given a new image, we generate sentences that are "nearby" the image in the same feature space. The encoder (mapping from image to feature space) is typically a Convolutional Neural Network (CNN), which abstracts images into a vector of local and global appearance features. The decoder (mapping from feature space to words) produces a word vector using a Recurrent Neural Network (RNN), which abstracts out the semantic and syntactic meaning.

For extracting visual features, Convolutional Neural Networks (CNNs) [53] have become very popular. A typical CNN is much like a classical feed-forward neural network, except that the connections between early layers are not all fully-connected, but instead have specially-designed structures with shared weights that encode operations like convolutions and spatial pooling across local image regions. Modern CNNs are also typically very deep, often with 20 or more layers. While the final output of CNNs differ based on the task (e.g., a class label for image classification), a common trick is to use the output of one of the penultimate layers as a feature vector to represent the visual appearance of an input image. These "deep features" produced by CNNs have been repeatedly shown [74,32] to outperform traditional hand-made image features such as SIFT [60] and HOG [18].

For modeling sequences like sentences, Recurrent Neural Networks (RNNs) [24,34], and specifically Long Short-Term Memory (LSTM) models [34,37], have become popular. RNNs include hidden units that are self-connected (i.e. some of their inputs are connected to their own outputs), which have the effect of "memory" that can develop internal representations for the patterns of input sequences [24]. LSTMs are a special form of RNNs that include an array of specially-designed hidden units called memory blocks, each of which contains three gates and a memory cell. In any given iteration of training, a memory block can choose to read or ignore its input, to remember or forget its current cell value, or to output or suppress the new cell value. These two ingredients of CNN and LSTM models are usually combined for image captioning in the following way [89,46,88]. The training data consists of a set of images, each with at least one human-generated reference sentence. During training, for any given image I, we first generate a corresponding deep visual feature vector v_I using a CNN. This vector is then presented as the initial input to the LSTM model. Then, the LSTM model is presented with each word in the training sentence in turn, by inputing the word vector corresponding to each word to the LSTM. The output from the hidden unit predicts the next word in the sentence, in

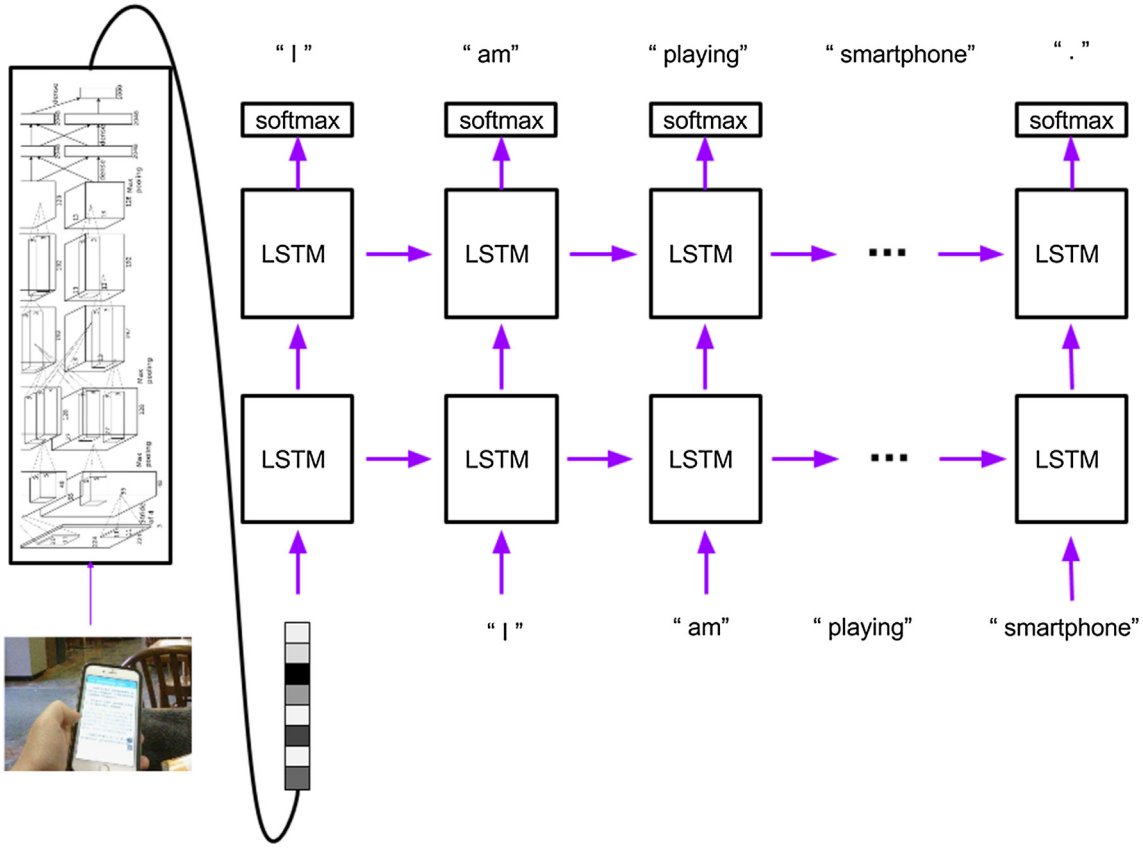


Fig. 5. *Illustration of deep image captioning using a CNN and a two-layer LSTM.* An image is fed into a CNN to produce a visual feature representation, and presented to the LSTM as its initial input. Then each word in a training sentence is presented to the LSTM at each step. A softmax layer is attached to the second layer of the LSTM to generate predicted sentences and softmax loss.

particular giving a probability distribution over all words in the dictionary. During each step t of training, error is back-propagated from the following step as well as from the softmax (word generation) layer. Word vectors as well as weights of hidden units are trained and updated during back-propagation. An intuitive way to visualize an LSTM is to unroll the hidden states at each time step, as shown in Fig. 5. In practice we use a two-layer LSTM to get better predictions, such that the hidden states of the first layer serve as input to the second layer and word predictions are emitted from the output of the second layer.

In testing, to generate a caption for new image I', we again use a CNN to produce an image feature vector and present it to the LSTM, which then predicts the first word of the sentence based on the visual features. After that, the best prediction at step t for word w_t is used as the input for step $t + 1$. In the prediction stage, a forward pass of the LSTM generates a full sentence terminated by a stop word for each input image. Similar image captioning models have been discussed in detail in recent papers [46,88,89]. We make several important improvements on this basic recipe to adapt captioning to the lifelogging image domain.

4.2. Photo grouping and activity summarization

The techniques in the last section automatically estimate captions for individual images. However, lifelogging users do not typically capture individual images in isolation, but instead collect long streams of photos taken at regular intervals over time (e.g., every 30 s for Narrative Clip). This is a significant difference from most applications of image captioning that have been studied before, which target isolated images found on the Internet or in social media, and represents both a challenge and an opportunity. The challenge is that any given individual lifelogging photo tends to be blurry, poorly composed, and otherwise much noisier than traditional photographs. Moreover, generating thousands of captions for a day's activities, one for each photo, could easily overwhelm a user. The opportunity is that evidence from multiple images can be combined to produce better captions than is possible from observing any single image, in effect "smoothing out" noise in any particular image by examining the photos taken nearby in time. These sentences could provide more concise summarizations, helping people find, remember, and organize photos according to broad events instead of individual moments. In fact, the sequence of these summarized sentences could automatically create a short textual "diary" of one's day!.

Suppose we wish to estimate captions for a stream of images $I = (I_1, I_2, ..., I_K)$, which are sorted in order of increasing timestamps. We first generate multiple diverse captions for each individual image, using a technique we describe in the next subsection. We combine all of these sentences across images into a large set of candidates C (with $|C| = d|I|$, where d is the number of diverse sentences generated per image; we use $d = 15$). We wish to estimate a sequence of sentences such that each sentence describes its corresponding image well, but also such that the sentences are relatively consistent across time. In other words, we want to estimate a sequence of sentences $S^* = (S_1^*, S_2^*, ..., S_K^*)$ so as to minimize an energy function,

$$S^* = \underset{S=(S_1,...,S_K)}{\operatorname{argmin}} \sum_{i=1}^{K} \text{Score}(S_i, I_i) + \beta \sum_{j=1}^{K-1} \mathbb{1}(\mathbb{S}_j, \mathbb{S}_{j+1}), \tag{1}$$

where each $S_i \in C$, $\text{Score}(S_i, I_i)$ is a unary cost function measuring the quality of a given sentence S_i in describing a single image I_i, $\mathbb{1}(\mathbb{S}_a, \mathbb{S}_b)$ is a pairwise cost function that is 0 if S_a and S_b are the same and 1 otherwise, and β is a constant. Intuitively, β controls the degree of temporal smoothing of the model: when $\beta = 0$, for example, the model

simply chooses sentences for each image independently without considering neighboring images in the stream, whereas when β is very large, the model will try to find a single sentence to describe all of the images in the stream.

Eq. (1) is a chain-structured Markov Random Field (MRF) model [51], which means that the optimal sequence of sentences S^* can be found efficiently using the Viterbi algorithm. All that remains is to define two key components of the model: (1) a technique for generating multiple, diverse candidate sentences for each image, in order to obtain the candidate sentence set C, and (2) the score function, which requires a technique for measuring how well a given sentence describes a given image. We now describe these two ingredients in turn.

4.2.1. Generating diverse captions

Our joint captioning model above requires a large set of candidate sentences. Many possible sentences can correctly describe a given image, and thus it is desirable for the automatic image captioning algorithm to generate multiple sentences that describe the image in multiple ways. This is especially true for lifelogging images that are often noisy, poorly composed, and ambiguous, and can be interpreted in different ways. Vinyals et al. [89] use beam search to generate multiple sentences, by having the LSTM model keep b candidate sentences at each step of sentence generation (where b is called the beam size). However, we found that this existing technique did not work well for lifelogging sentences, because it produced very homogeneous sentences, even with a high beam size.

To encourage greater diversity, we apply the diverse m-best solutions technique of Batra et al. [8], which was originally proposed to find multiple high-likelihood solutions in graphical model inference problems. The idea is to perform inference multiple times, each time finding the optimal solution that is sufficiently different from all previously found solutions, which can be computed efficiently by modifying the unary cost function in each iteration based on prior solutions. We adapt this technique to LSTMs by performing multiple rounds of beam search. In the first round, we obtain a set of predicted words for each position in the sentence. In the second round, we add a bias term that reduces the network activation values of words found in the first beam search by a constant value. Intuitively, this decreases the probability that a word found during the previous beam search will be selected again at the same word position in the sentence. Depending on the degree of diversity needed, additional rounds of beam search can be conducted, each time penalizing words that have occurred in any previous round. In our current implementation, we use three rounds of beam search and set the beam size to be five, so we generate a total of 15 candidate sentences for each individual image. The set of all of these sentences across all images in the photo stream produce the candidate sentence set C in Eq. (1).

Fig. 6 presents sample automatically-generated results using three rounds of beam search and a beam size of 3 for illustration purposes. We see that the technique successfully injects diversity into the set of estimated captions. Many of the captions are quite accurate, including "A man is sitting at a table" and "I am having dinner with my friends," while others are not correct (e.g. "A man is looking at a man in a red shirt"), and others are nonsensical ("There is a man sitting across the table with a man"). Nevertheless, the captioning results are overall remarkably accurate for an automatic image captioning system, reflecting the power of deep captioning techniques to successfully model both image content and sentence generation.

4.2.2. Image-sentence quality alignment

The joint captioning model in Eq. (1) also requires a function called $\text{Score}(S_i, I_i)$, which measures how well an arbitrary sentence S_i describes a given image I_i. The difficulty here is that the LSTM model described above tells us how to generate sentences for an image, but not how to measure their similarity to a given image. Doing this requires us to explicitly align certain words of the sentence to certain regions of an image – i.e. determining which "part" of an image generated each word.

Karpathy et al. [48] propose matching each region with the word having the maximum inner product (interpreted as a similarity measure) across all words in terms of learnable region vectors and word vectors, and then summing all similarity measures over all regions as the total score. In their approach, image regions are detected and region features are generated by R-CNNs [32], a popular deep learning-based object detector. Word vectors are encoded with Bidirectional LSTMs (BLSTM) [35,78], which are a variant of LSTMs that capture contextual information not only from previous words but also from future ones. We follow their approach and train this image-sentence alignment model on our lifelogging dataset. To generate the matching score $\text{Score}(S_i, I_i)$ for Eq. (1), we extract region vectors from image I_i, retrieve trained word vectors for words in sentence S_i, and sum the similarity measures of the regions with the best-aligned words. This provides a natural image-sentence matching score to quantify the similarity of a caption with an image.

Here we give a brief review of their method (albeit with different notation to clarify details). Suppose for an image I, there are M region descriptors $B_I = \{b_1, ..., b_M\}$ for bounding boxes detected by the R-CNN, and N word vectors $D_I = \{d_1, ..., d_N\}$ for words in its description sentence. We first define a true image-sentence pair matching score $S_I = P_{I,I} + Q_{I,I}$, with

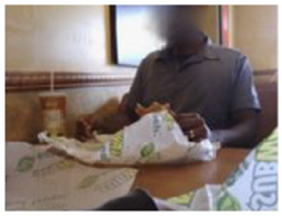
$$P_{I,I} = \sum_{b_i \in B_I} \max_{d_j \in D_I} d_j^T b_i, \text{ and}$$
$$Q_{I,I} = \sum_{d_i \in D_I} \max_{b_j \in B_I} b_j^T d_i. \quad (2)$$

Each region is matched with a single word of maximum inner product and accumulates this value for all regions as $P_{I,I}$, while each word is matched with a single region of maximum inner product and accumulates this value for all words as $Q_{I,I}$. $P_{I,I}$ measures how regions could be best described by words, and $Q_{I,I}$ measures how words could be best explained by part of the image. The sum of $P_{I,I}$ and $Q_{I,I}$ gives the matching score of image and sentence.

4.2.3. Image grouping

Finally, once captions have been jointly inferred for each image in a photostream, we can group together contiguous substreams of images that share the same sentence. Fig. 7 shows how we use the Viterbi algorithm to search for optimal solutions of Eq. (1). We first generate a pool of diverse candidate sentences for the images (for ease of visualization, only three are shown here), each of which has a prior probability (blue column). For each image and each sentence, we calculate the Score function (yellow matrix). The Viterbi algorithm run on the MRF can be thought of as finding the best "path" through the columns of this matrix to maximize the sum of the Scores while minimizing the number of sentence transitions (changes in rows) that are encountered. Two runs are shown, for (a) a large temporal smoothing weight β to force fewer sentence transitions (in this case, a single sentence), and (b) a small β that emphasizes the quality of match between each sentence and image, even if more sentence transitions must be used (causing the last sentence to be selected for the first four images but the first sentence to be selected for the final sentence).

Fig. 8 shows more examples of activity summarization. In general, the jointly-inferred captions are reasonable descriptions of the images, and much less noisy than those produced from individual images in Fig. 6, showing the advantage of incorporating temporal reasoning into the captioning process. For example, the last row of images shows that the model labeled several images as "I am talking with a friend while eating a meal in a restaurant," even though the friend is only visible in one of the frames, showing how the model has propagated context across time. Of course, there are still mistakes, ranging from the minor error that there is no broccoli on the plate in the first row to the more major error that the second to the last row shows a piano and not



Beam 1
1. a man is sitting at a table
2. i am having a dinner with my friends
3. i am having a dinner with a friend

Beam 2
1. a man is sitting by side of a table
2. a man is sitting at a table
3. a man is looking at a man in a red shirt

Beam 3
1. there is a man with glasses on the table
2. there is a man sitting across the table
3. there is a man across the table with a man

Beam 1
1. i am typing on my computer
2. i am meeting with my friend
3. a person is typing on a laptop

Beam 2
1. there is a computer monitor on the table
2. there is a computer monitor in the room
3. two hands are typing on a computer

Beam 3
1. I am typing on my computer
2. i am working on my computer
3. i am sitting with my friend

Beam 1
1. i am eating scrambled eggs with pepper
2. some scrambled eggs are sitting on a blue bowl
3. some watermelon is in a yellow bowl

Beam 2
1. i am eating breakfast of eggs with pepper
2. i am eating biscuits and snacks in a plate
3. a bowl of fruit is sitting on a blue bowl

Beam 3
1. there are some eggs sitting on a blue bowl
2. a breakfast with eggs and cakes
3. some eggs and cakes are sitting on a bowl

Beam 1
1. i am ordering food at a restaurant
2. i am ordering food in a restaurant
3. a man is preparing food in a restaurant

Beam 2
1. a woman and woman are preparing food in a restaurant
2. a woman is preparing food in a restaurant
3. a woman is preparing food in a cafeteria

Beam 3
1. i am ordering my food at a restaurant
2. i am talking to my friend at a restaurant
3. there is a man and woman at the table

Beam 1
1. i am driving underneath a blue sky
2. i am driving in a sunny day
3. i am driving underneath a blue sky with clouds

Beam 2
1. a blue sky and clouds are visible outside the train
2. a blue sky and clouds are visible through a windshield of a car
3. a blue sky and clouds are visible in view of a train

Beam 3
1. there are traffic lights and trees are underneath a blue sky
2. i am driving a car with clouds on my road
3. there is a traffic light in view of a train

Beam 1
1. i am seated at a hall
2. a group of people are sitting on a table in a cafeteria
3. a group of people are sitting around a table in a cafeteria

Beam 2
1. i am seated at a hall
2. i am seated at a restaurant
3. some people are sitting at a table

Beam 3
1. there are two friends at a restaurant
2. some people are sitting at a table
3. there are two hands in a cafeteria

Fig. 6. *Sample captions by models pre-trained with COCO and fine-tuned with Lifelog data,* showing the three highest-likelihood sentences for each of three beam searches.

someone typing on a computer. The grammar of the sentences is generally good considering that the model has no explicit knowledge of English besides what it has learned from training data, although usage errors are common (e.g., "I am shopping kitchen devices in a store"), some of which are created by the network and some of which can be traced back to grammar errors in the human-generated training sentences.

5. Experimental evaluation

We evaluate the various parts of our lifelogging image captioning approach under several different use cases and scenarios, in order to gain a better understanding of its strengths and weaknesses. A challenge in developing image captioning algorithms is how to evaluate them: a given image could be correctly described by a nearly boundless set of possible sentences, so simply checking whether a predicted sentence matches a ground truth sentences is woefully ineffective. Here we

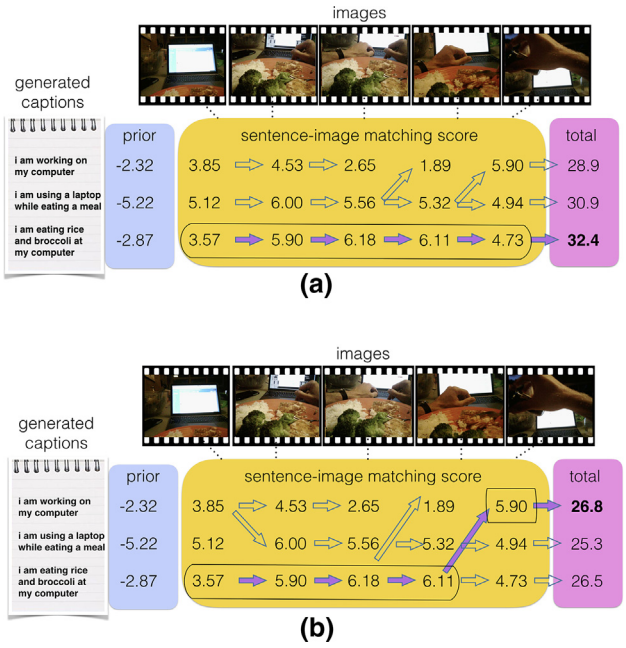


Fig. 7. Illustration of selecting summarized captions for a sequence of five images through our MRF-based formulation.

first use automatic metrics that compare to ground truth reference sentences; these are convenient because they produce quantitative scores, but are difficult to interpret. To give a better idea of the actual practical utility of the technique, we also evaluate it in two other ways: (1) by using a panel of human judges to rate the quality of captioning results, and (2) by testing the system in a specific application of keyword-based image retrieval using the generated captions.

5.1. Quantitative captioning evaluation

Automated metrics such as BLEU [71], CIDEr [87], Meteor [6] and Rouge-L [58] have been proposed to score sentence similarity compared to reference sentences provided by humans, and each has different advantages and disadvantages. BLEU was originally intended for evaluating machine translations but is also commonly used to evaluate captioning. BLEU counts occurrences of n-grams in a candidate sentence clipped by maximum occurrences in the reference sentences, and normalized by the total number of n-grams in the candidate. It is typically evaluated for different values of n, and in particular BLEU-n refers to the geometric mean from 1-grams to n-grams; here we report scores of BLEU-1 to 4. CIDEr [87] computes average cosine similarity (with TF-IDF weighting) between the n-grams (typically up to 4-grams) of the generated sentence and human-generated reference sentences. Meteor [6] compares unigrams between generated and reference sentences using different degrees of similarity (exact match, matching stems, synonymy). Rouge-L [58] produces an F-measure based on length of longest common sequence of candidate and reference sequence. We present results using all of these metrics (using the COCO Detection Challenge implementation[4]), and also summarize the seven scores with their means.

5.1.1. Implementation

A significant challenge with deep learning-based methods is that they typically require huge amounts of training data, both in terms of number of images and number of sentences; for example, the winner of the ImageNet classification challenge [53] trained on 15 million images, while for image captioning, Vinyals et al. [89] trained on COCO [59], which has over 80,000 photos and 400,000 sentences. Unfortunately, collecting this quantity of lifelogging images and annotations is very difficult, compared to the COCO images which were downloaded from public Internet sources like Flickr. To try to overcome this problem, we augmented our lifelogging training set with COCO data using three different strategies: **Lifelog only** training used only our lifelogging dataset, consisting of 736 lifelogging photos with 4,300 human-labeled sentences; **COCO only** training used only COCO dataset; and **COCO then Lifelog** started with the **COCO only** model, and then used it as initialization when re-training the model on the lifelogging dataset (a strategy often referred to as "fine-tuning" in the deep learning literature [53]).

For extracting image features, we use the VGGNet [80] CNN model since it was found to work best on this problem by Karpathy et al. [46], and because a Caffe implementation was readily available [43]. The word vectors are learned from scratch. Our image captioning model stacks two LSTM layers, and each layer structure closely follows the one described in [89]. For training the LSTM module, we use a two-layer LSTM with hidden size 256, and set the learning rate to 0.0004, the decay rate to 0.997, the batch size to 128, and use 10,000 epochs in total. To boost training speed, we implemented the LSTM model in C++ using the Caffe [43] deep learning package (as opposed to relying on the publicly-available Python implementation, which is orders of magnitude slower; we have made our source code available online[5]). Our implementation requires about 2.5 h for COCO pre-training with 10,000 iterations, and about 1 h for fine-tuning on the Lifelog dataset with 10,000 iterations on a dual-processor Xeon 2.5 GHz system with two NVidia K40 GPUs.

In testing, the number of beam searches conducted during caption inference controls the degree of diversity in the output; here we use three to match the three styles of captions we expect (COCO, first-person, and third-person perspectives). Samples of predicted sentences are shown in Fig. 6. This suggests that different genres of training sentences contribute to tuning the hidden states of the LSTM and thus enable it to produce diverse structures of sentences in the testing stage. Our unoptimized captioning generation process takes about 0.5 s per image on the same system described above.

5.1.2. Results

Fig. 9 plots detailed quantitative results of each of these training strategies, in terms of the Bleu, CIDEr, Meteor, and Rouge scores. The left three plots of the graph are all tested on the same set of 100 randomly-selected photos having 1,000 ground truth reference sentences. We find that the **Lifelog only** strategy achieves much higher overall accuracy than **COCO only**, with a mean score (across all seven metrics) of 0.373 vs. 0.272. This suggests that even though COCO is a much larger dataset, images from social media are different enough from lifelogging images that the **COCO only** model does not generalize well to our application. Moreover, this may also reflect an artifact of the automated evaluation, because **Lifelog only** benefits from seeing sentences with similar vocabulary and in a similar style as in the reference sentences, since the same small group of humans labeled both the training and test datasets. More surprisingly, we find that **Lifelog only** also slightly outperforms **COCO then Lifelog** (0.373 vs 0.369). The model produced by the latter training dataset has a larger vocabulary and produces richer styles of sentences than **Lifelog only**, which hurts its quantitative score. Qualitatively, however, it often produces more diverse and descriptive sentences because of its larger vocabulary and ability to generate sentences in first-person, third-person, and COCO styles. Fig. 10 compares captions produced by two of these models for some sample images, while samples of generated diverse captions are shown in Fig. 6.

[4] https://github.com/tylin/coco-caption.

[5] http://vision.soic.indiana.edu/deepdiary.

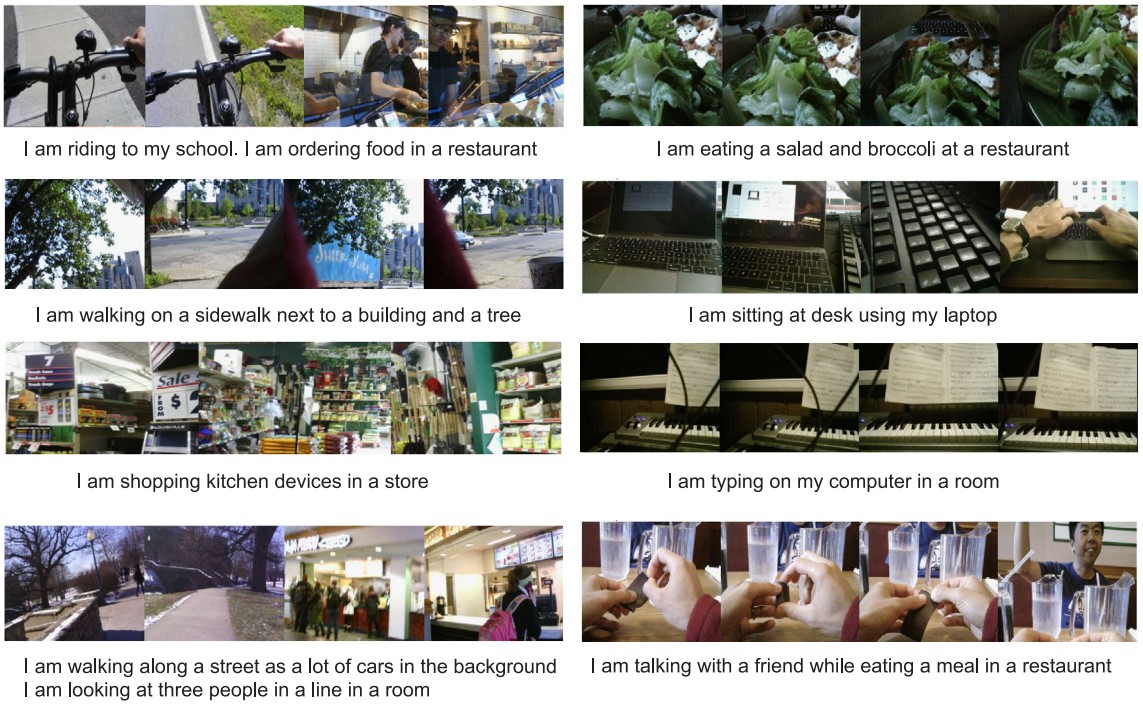


Fig. 8. Randomly-chosen samples of activity summarization on our dataset.

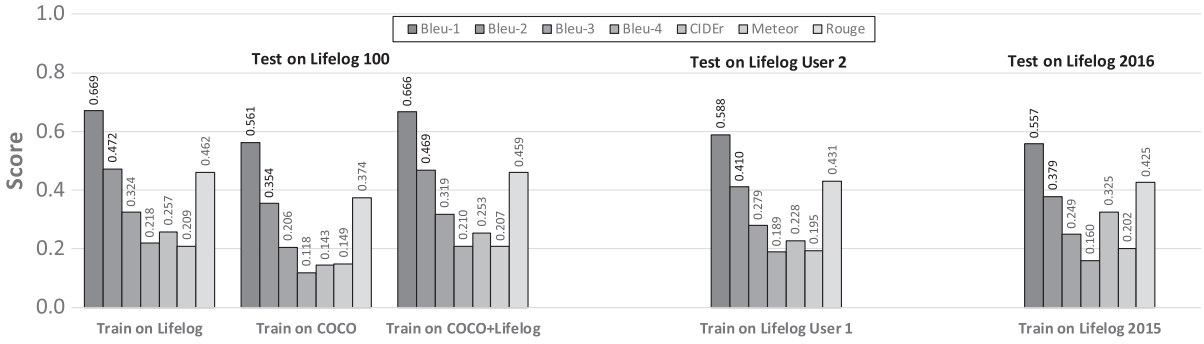


Fig. 9. *Quantitative evaluation of captioning results,* in terms of Bleu1-4, CIDEr, Meteor and Rouge scores for Diverse 3-Best beams of captions, for various combinations of testing and training data.

Our experiments so far use separate training and test sets, of course, but both draw images from the same pool of lifelogging images. This means that very similar photos may appear in the two sets, since a photo taken at one moment may be nearly identical to one taken in the next. We conducted experiments with two additional strategies in order to simulate more realistic scenarios. The first scenario reflects when a consumer first starts using our automatic captioning system on their images without having supplied any training data of their own. We simulate this by training the image captioning model on one user's photos and testing on another. This training set has 805 photos and 3716 reference sentences while the testing set has 40 photos and 565 reference sentences. The mean quantitative accuracy declines from our earlier experiments when training and testing on images sampled from the same set, as shown in the middle plot of Fig. 9, although the decline is not very dramatic (from 0.373 to 0.331), and still much better than training on COCO (0.272). This result suggests that the captioning model has learned general properties of lifelogging images, instead of overfitting to one particular user (e.g., simply "memorizing" the appearance of the places and activities they frequently visit and do).

Another realistic situation is when an existing model trained on historical lifelogging data is used to caption new photos. We simulate this by taking all lifelogging photos in 2015 as training data and photos in 2016 as testing data; here the training set has 673 photos and 3,610 sentences and the testing set has 30 photos and 172 sentences. As shown in the right plot of Fig. 9, this scenario very slightly decreased performance compared to training on data from a different user (0.328 vs 0.331).

5.2. Image captioning evaluation with human judges

The evaluation metrics used in the last section are convenient because they can be automatically computed from ground-truth reference sentences, and are helpful for objectively comparing different methods. However, they give little insight into how accurate or descriptive the sentences are, or whether they would be useful for real lifelogging users.

We conducted a small study using human judges to rate the quality of our automatically-generated captions. In particular, we randomly selected 21 images from the Lifelog 100 test dataset (used in Fig. 9) and generated captions using our model trained on the COCO then Lifelog scenario. For each image, we generated 15 captions (with 3 rounds of beam search, each with beam size 5), and then kept the top-scoring

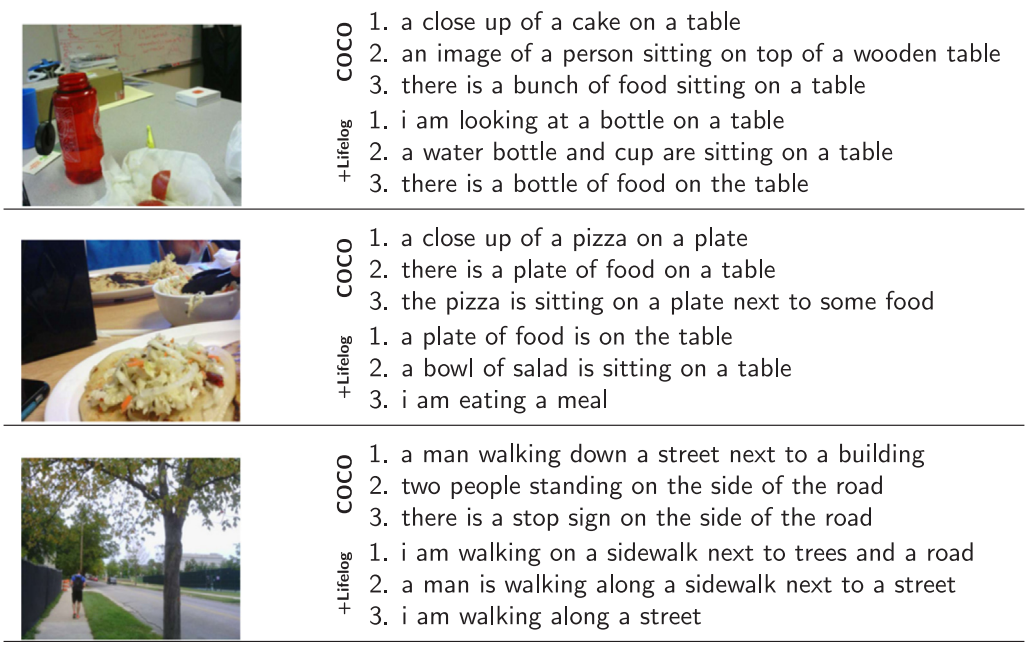


Fig. 10. Image captions produced by models trained on COCO, without and with fine-tuning on LifeLog.

caption according to our model and four randomly-sampled from the remaining 14, to produce a diverse set of five automatically-generated captions per image. We also randomly sampled five of the human-generated reference sentences for each image.

For each of the ten captions (five automatic plus five human), we showed the image (after reviewing it for potentially private content and obtaining permission of the photo-taker) and caption to users on Amazon Mechanical Turk, without telling them how the caption had been produced. We asked them to rate, on a five-point Likert scale, how strongly they agreed with two statements: (1) "The sentence or phrase makes sense and is grammatically correct (ignoring minor problems like capitalization and punctuation)," and (2) "The sentence or phrase accurately describes either what the camera wearer was doing or what he or she was looking at when the photo was taken." We also included a field for free-form comments. We once again required United States residency and a Mechanical Turk track record of at least 100 HITs and at least a 97% approval rating. Each image and 10 associated captions were shown to two Turk users, for a total of 420 individual HITs and 37 users. In a pilot study we estimated that the time required to complete a HIT (judging a single sentence according to the two statements above) took about 15 s, so we paid 5 cents per HIT (corresponding to an equivalent hourly wage of about US$12). Inter-rater agreement was high: for the first question, the two Mechanical Turk users selected the same five-point Likert option for 79.0% of the captions and were within 1 option of one another for 93.5% of the captions, and for the second question, the users exactly agreed for 75.5% of the captions and agreed within 1 point for 90.5% of them.

Fig. 11 summarizes the results, comparing the average ratings over the 5 human reference sentences, the average over all 5 diverse automatically-generated captions (Auto-5 column), and the single highest-likelihood caption as estimated by our complete model (Auto-top). About 92% of the human reference sentences were judged as grammatically correct (i.e., somewhat or strongly agreeing with statement (1)), compared to about 77% for the automatically-generated diverse captions and 81% for the single best sentence selected by our model. Humans also described images more accurately than the diverse captions (88% vs 54%), although the fact that 64% of our single best estimated captions were accurate indicates that our model is often able to identify which one is best among the diverse candidates. Overall, our top automatic caption was judged to be *both* grammatically correct and

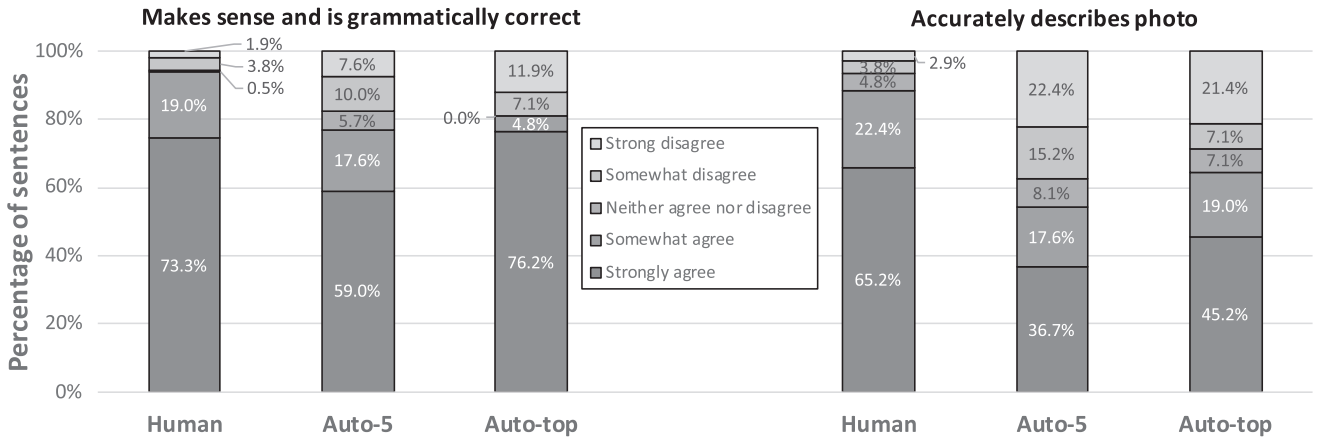


Fig. 11. *Summary of grammatical correctness and accuracy of lifelogging image captions,* on a rating scale from 1 (Strongly Disagree) to 5 (Strongly Agree), averaged over 3 human judges. *Human* column is averaged over 5 human-generated reference sentences, *Auto-5* is averaged over 5 diverse computer-generated sentences, and *Auto-top* is single highest-likelihood computer-generated sentence predicted by our model.

Table 1
Confusion matrices for two approaches on two tasks involving detecting sensitive images. *Top:* Results on 3-way problem of classifying into not sensitive, sensitive place (bathroom), or digital display categories. *Bottom:* Results on 2-way problem of classifying into sensitive or not (regardless of sensitivity type). Actual classes are in rows and predicted classes are in columns.

	CNN classifier			Caption classifier		
	Not sens	Place	Display	Not sens	Place	Display
Not sensitive	0.730	0.130	0.140	0.686	0.117	0.197
Sensitive place	0.189	0.811	0	0.151	0.792	0.057
Display	0.300	0.043	0.657	0.143	0.008	0.849

	Not sens	Sensitive	Not sens	Sensitive
Not sensitive	0.730	0.270	0.686	0.314
Sensitive	0.317	0.683	0.161	0.839

accurate 59.5% of the time, compared to 84.8% of the time for human reference sentences. Using an ordinal metric in which 1 is Strong Disagree and 5 is Strong Agree, the mean grammar score across all judges and sentences was 4.26 for the most likely automatic sentence, compared to 4.60 for the human-generated sentences, and the mean accuracy score was 3.60 for the automatic sentences compared to 4.45 for the human-generated sentences.

We view these results to be very promising, as they suggest that automatic captioning can generate reasonable sentences for over half of lifelogging images, at least in some applications. For example, for 19 (90%) of the 21 images in the test set, at least one of the five diverse captions was unanimously judged to be both grammatically correct and accurate by all 3 judges. This may be useful in some retrieval applications where recall is important, for example, where having noise in some captions may be tolerable as long as at least one of them is correct. We consider one such application in the next section.

5.3. Keyword-based image retrieval

Image captioning allows us to directly implement keyword-based image retrieval by searching on the generated captions. We consider a particular application of this image search feature here that permits a quantitative evaluation. As mentioned above, wearable cameras can collect a large number of images containing private information. Automatic image captioning could allow users to find potentially private images easily, and then take appropriate action (like deleting or encrypting the photos). We consider two specific types of potentially embarrassing content here: photos taken in potentially private locations like bathrooms and locker rooms, and photos containing personal computer or smartphone displays which may contain private information such as credit card numbers or e-mail contents.

We chose these two types of concerns specifically because they have been considered by others in prior work: Korayem et al. [52] present a system for detecting monitors in lifelogging images using deep learning with CNNs, while Templeman et al. [86] classify images according to the room in which they were taken. Both of these papers present strongly supervised based techniques, which were given thousands of training images manually labeled with ground truth for each particular task. In contrast, identifying private imagery based on keyword search on automatically-generated captions could avoid the need to create a training set and train a separate classifier for each type of sensitive image.

We evaluated captioning-based sensitive image retrieval against standard state-of-the-art strongly-supervised image classification using CNNs [53] (although we cannot compare directly to the results presented in [52] or [86] because we use different datasets). We trained the strongly-supervised model by first generating a training set consisting of photos having monitors and not having monitors, and photos taken in bathrooms and locker rooms or elsewhere, by using the ground truth categories given in the COCO and Flickr8k datasets. This yielded 34,736 non-sensitive images, 6,135 images taken in sensitive places, and 4379 images with displays. We used a pre-trained AlexNet model (1000-way classifier on ImageNet data) fine-tuned on our dataset by replacing the final fully connected layer with a 3-way classifier to correspond with our three-class problem.

We also ran the technique proposed here, where we first generate automatic image captions, and then search through the top five captions for each image for a set of pre-defined keywords (specifically "toilet," "bathroom," "locker," "lavatory," and "washroom" for sensitive place detection, and "computer," "laptop," "iphone," "smartphone," and "screen" for display detection). If any of these keywords is detected in any of the five captions, the image is classified as sensitive.

Table 1 presents the confusion matrix for each method, using a set of 600 manually-annotated images from our lifelogging dataset as test data (with 300 non-sensitive images, 53 images in sensitive places, and 252 with digital displays). We see the supervised classifier has better prediction performance on finding sensitive places (0.811) than the caption-based classifiers (0.792), while the caption-based classifier outperforms the CNN on predicting the second type of sensitive image: displays (0.849 vs 0.657). In a real application, determining the type of private image is likely less important than simply deciding if it is private. The bottom row of Table 1 reflects this scenario, showing a confusion matrix which combines the two sensitive types and focuses on whether photos are sensitive or not.

From another point of view, sensitive photo detection is a retrieval problem. Fig. 12 shows precision-recall curves for CNN and caption-based classifiers. They show the trade-off between selecting accurate sensitive photos (high precision) and obtaining a majority of all sensitive photos (high recall). For example, by using the CNN classifier, we can obtain 80% of type 1 (sensitive place) photos with accuracy around 58% (Fig. 12(left), green curve); by using the caption-based classifier, we can obtain 80% of type 2 (digital display) sensitive photos with precision around 78% (Fig. 12(right), blue curve). Overall, these results suggest that keyword search in automatically-generated captions could yield similar accuracies to strongly-supervised classifiers, but without explicit re-training on each type of private image. The two approaches may also be complementary, since they use different forms of evidence in making classification decisions, and users in a real application could choose their own trade-off on how aggressively to filter lifelogging images.

6. Discussion

We have proposed the idea of using automatic image captioning in

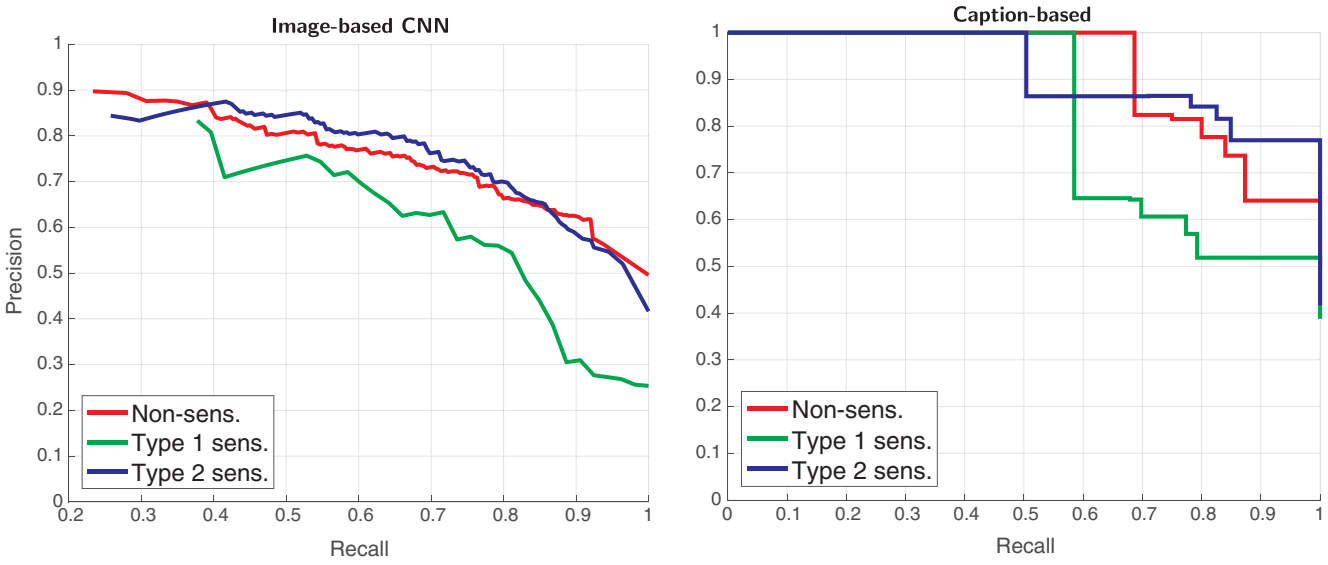


Fig. 12. *Precision-recall curves for retrieving sensitive images* using CNNs (left) and generated captions (right).

order to help users summarize and organize their lifelogging photos. Our experimental results are promising, both for image captioning itself and for derived tasks like retrieval based on keyword search on the generated captions. Our experiments simulated several use cases including when training and testing photos are taken by different users, or when they are taken at different times.

However, we stress that these are preliminary experiments with a number of limitations. Our lifelogging dataset was collected by only two people, both of whom are computer scientists living in the same town. Less constrained situations are likely to be more challenging for the captioning system. On the other hand, our performance was likely limited by the small amount of training data we could collect, while a deployed system could create much larger training datasets from the photos and reference sentences of its users. Of course, people's lives have a large degree of structure and regularity, which could also be exploited in a real system. For example, a system could use metadata like timestamps and GPS to identify when and where a photo was taken, and match that metadata to external data sources (like GIS maps) to produce more accurate captioning results. Moreover, the regularity itself could be exploited: even if there are no reference sentences available for a particular user, the system could use training data from other similar users to create more personalized captions. A real-world system could train multiple models for use with different demographic groups (e.g., based on age, occupation, gender, location, culture, etc.). Meanwhile, some details – the names of the people that appear in a user's lifelogs, for example – are so specific to a single user that no degree of pre-training could accurately identify them. A real system could give some mechanism for users to write their own reference sentences for some of their photos, which the system could use as training data in order to customize to a particular user's environment and lifestyle.

To give some idea of how a model trained on one first-person dataset would perform on a completely unrelated dataset, Fig. 13 presents sample captions on the Dogcentric Activity Dataset [42], but trained on COCO and our Lifelog datasets. The Dogcentric dataset was taken from a first-person camera mounted on a dog, and thus the types of interactions and perspectives it captured were completely different from our data. Nevertheless, our results seem reasonable, including identifying key objects (buildings, umbrellas, fields), even if the exact details are not quite right. We see that the model trained with COCO inherits that dataset's third-person captioning style, and produces captions from a bystander's point of view, while the model trained with Lifelog tends to describe scenes from first-person viewpoints. Future work could try to rigorously characterize the relationship between size and diversity of training set and the performance in less-constrained use scenarios.

We have considered the problem of writing captions for an entire image, but real applications may also benefit from models that map parts of a caption to specific parts of the image. This could be useful to explain the visual evidence used by the captioning algorithm, for example, which may in turn be useful for "debugging" unexpected captioning results. As a first step, we adapted the method of Selvaraju et al. [79] to illustrate the connection between image parts and predicted words. For a given word of interest predicted by the LSTM, we flow the gradient of its class label back through all its previous steps as well as through the entire VGG network until it reaches the very last convolutional layer. These gradients are global-average-pooled to serve as weights of those feature maps obtained in the last convolutional layer. Finally a weighted combination is computed over all those feature maps and a ReLU is applied to produce a final "attention map" for that word. The attention map is resized to the dimensions of the original image, and overlaid as a heat map. We show some sample heat maps in Fig. 14, where red indicates the strongest connection to the word of interest. These results are intuitive: "driving" is associated with the steering wheel and hands in the first image, whereas "laptop" is associated with the computer in the second image. In the third image, the visualizations help to explain the mention of "fireworks:" the network has incorrectly identified the colored lights of the building at night as fireworks against a night sky.

The experiments in this paper were conducted on a high-end workstation with an advanced GPU (an NVidia Tesla K40) to make deep learning training practical. Most end users do not have this hardware available, so we envision the system being deployed as a cloud-based service. Existing lifelogging platforms including Narrative require users to upload photos to the cloud anyway, so we do not anticipate this to be a major barrier.

7. Conclusion

Unlike other types of lifelogging that have become mainstream, camera-based lifelogging is still in its infancy: the practical hardware issues involved with creating low-cost, efficient, wearable cameras seem to have been largely solved, but how to manage these photos is still a significant challenge, and one that very well may determine whether or not consumers actually adopt the technology. We have proposed the idea of automatically generating captions to "narrate" lifelogging photo streams, as a method to help people summarize their

Fig. 13. *Captions automatically generated for sample frames from the Dogcentric dataset,* comparing models trained on COCO and Lifelog datasets.

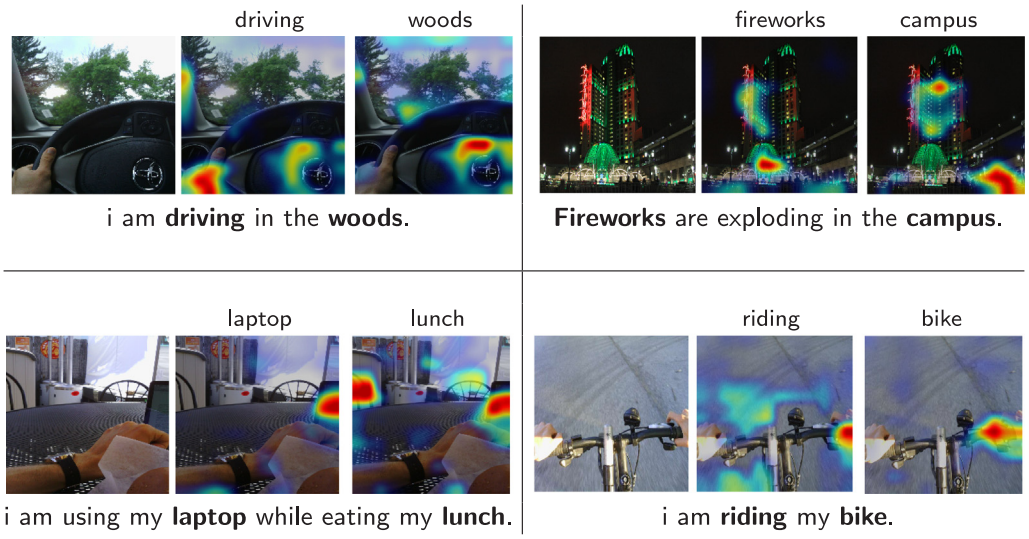


Fig. 14. Visualizations of connections between image evidence and predicted words.

daily activities, find and recall photos that they may be interested in keeping, and removing photos that may want to delete due to private or sensitive content. We used existing deep captioning models with two novel techniques to better suit the lifelogging context: (1) jointly captioning streams of photos instead of individual photos in isolation using a Markov Random Field formulation, and (2) injecting diversity into captions to help address photos that do not have a single obvious subject (as is common in lifelogging images). We have also introduced and released a large-scale lifelogging dataset with image caption annotations.

Our work is a first step towards exploring captioning for lifelogging images. Our experiments suggest that while the results are not perfect, they could already work well enough to help, and will undoubtedly continue to improve as the novel area of captioning lifelogging images receives more attention in the future. An important component of practical systems will be to use human feedback to personalize captioning to one particular user's environment and preferences [72]. Another interesting future direction is to expand the work beyond captioning and into visual question answering [2,33,19], which could answer a user's specific questions about their lifelogging stream, or produce captions from one particular desired perspective.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

This work was supported in part by the National Science Foundation (CAREER IIS-1253549 and CNS-1408730) and Google, and the IU Office of the Vice Provost for Research, the College of Arts and Sciences, and the School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project, "Learning: Brains, Machines, and Children." CF was supported by a Paul Purdom Fellowship. This work used compute facilities provided by NVidia, the Lilly Endowment through support of the Indiana University Pervasive Technology Institute, the Indiana METACyt Initiative, and the Romeo FutureSystems facility which is partially supported by Indiana University and NSF RaPyDLI grant 1439007. We thank Zhenhua Chen, Sally Crandall, and Xuan Dong for helping to label our lifelogging photos, and Katherine Spoon for copy-editing corrections and suggestions.

References

[1] T. Ahmed, P. Shaffer, K. Connelly, D. Crandall, A. Kapadia, Addressing physical safety, security, and privacy for people with visual impairments, in: USENIX Symposium on Usable Privacy and Security (SOUPS) (2016).
[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and vqa. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018).
[3] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural module networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) pp. 39–48.
[4] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, B. MacIntyre, Recent advances in augmented reality, IEEE Comput. Graphics Appl. 21 (6) (2001) 34–47.
[5] S. Bambach, S. Lee, D. Crandall, C. Yu, Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions, in: IEEE International Conference on Computer Vision (ICCV) (2015).
[6] S. Banerjee, A. Lavie, Meteor: an automatic metric for MT evaluation with improved correlation with human judgments. Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (2005) 65.
[7] D. Barreau, A. Crystal, J. Greenberg, A. Sharma, M. Conway, J. Oberlin, M. Shoffner, S. Seiberling, Augmenting memory for student learning: designing a context-aware capture system for biology education, Am. Soc. Inf. Sci. Technol. 43 (1) (2006) 1–6.
[8] D. Batra, P. Yadollahpour, A. Guzman-Rivera, G. Shakhnarovich, Diverse m-best solutions in markov random fields, European Conference on Computer Vision (ECCV), Springer, 2012, pp. 1–16.
[9] BBC News, September 26, 2016. Shutter falls on life-logging camera start-up Narrative. BBC News.
[10] A. Betancourt, P. Morerio, C. Regazzoni, M. Rauterberg, The evolution of first person vision methods: a survey, IEEE T. Circuits Syst. Video Technol. 25 (5) (2015) 744–760.
[11] M. Bolanos, M. Dimiccoli, P. Radeva, Towards storytelling from visual lifelogging: an overview. arXiv preprint 1507.06120 (2015).
[12] M. Bolanos, A. Peris, F. Casacuberta, S. Soler, P. Radeva, Egocentric video description based on temporally-linked sequences, J. Visual Commun. Image Represent. 50 (2018) 205–216 URL http://www.sciencedirect.com/science/article/pii/S1047320317302316.
[13] L.A. Cadmus-Bertram, B.H. Marcus, R.E. Patterson, B.A. Parker, B.L. Morey, Randomized trial of a fitbit-based physical activity intervention for women, Am. J. Preventive Med. 49 (3) (2015) 414–418.
[14] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, I. Essa, Predicting daily activities from egocentric images using deep learning, in: Intl. Symposium on Wearable Computers (2015).
[15] X. Chen, C.L. Zitnick, Mind's eye: A recurrent visual representation for image caption generation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) pp. 2422–2431.
[16] S. Clinch, P. Metzger, N. Davies, Lifelogging for observer view memories: an infrastructure approach, in: 2014 ACM Intl. Joint Conf. on Pervasive and Ubiquitous Computing: Adjunct Publication. (2014) pp. 1397–1404.
[17] D. Crandall, C. Fan, Deepdiary: Automatically captioning lifelogging image streams, in: European Conference on Computer Vision International Workshop on Egocentric Perception, Interaction, and Computing (EPIC) (2016).
[18] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conf. Comput. Vision Pattern Recogn. (CVPR). 1 (2005) 886–893.
[19] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, D. Batra, Embodied question answering. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018).
[20] T. Denning, Z. Dehlawi, T. Kohno, In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies, in: ACM CHI Conference on Human Factors in Computing Systems (CHI) (2014) pp. 2377–2386.
[21] A. Doherty, K. Pauly-Takacs, N. Caprani, C. Gurrin, C. Moulin, N. O'Connor, A. Smeaton, Experiences of aiding autobiographical memory using the SenseCam, Human-Comput. Interact. 27 (1–2) (2012) 151–174.
[22] A.R. Doherty, N. Caprani, V. Kalnikaite, C. Gurrin, A.F. Smeaton, E. Noel, et al., Passively recognising human activities through lifelogging, Comput. Hum. Behav. 27 (5) (2011) 1948–1958.
[23] J. Donahue, L.A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 39 (4) (2017) 677–691. URL doi: http://dx.doi.org/10.1109/TPAMI.2016.2599174.
[24] J.L. Elman, Finding structure in time, Cognitive Sci. 14 (2) (1990) 179–211.
[25] Engadget, 2013. Narrative clip. https://www.engadget.com/products/narrative/clip/.
[26] D. Erhan, C. Szegedy, A. Toshev, D. Anguelov, Scalable object detection using deep neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) pp. 2155–2162.
[27] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J.C. Platt, et al., From captions to visual concepts and back, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) pp. 1473–1482.
[28] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: generating sentences from images, in: K. Daniilidis, P. Maragos, N. Paragios (Eds.), European Conference on Computer Vision (ECCV), Springer, Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 15–29.
[29] A. Fathi, Y. Li, J.M. Rehg, Learning to recognize daily actions using gaze, European Conference on Computer Vision (ECCV), Springer, 2012, pp. 314–327.
[30] A. Fathi, X. Ren, J.M. Rehg, Learning to recognize objects in egocentric activities, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011) pp. 3281–3288.
[31] A. Furnari, G. Farinella, S. Battiano, Recognizing personal contexts from egocentric images, in: ICCV Workshops (2015).
[32] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) pp. 580–587.
[33] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017).
[34] A. Graves, Generating sequences with recurrent neural networks. arXiv:1308.0850 (2013).
[35] A. Graves, N. Jaitly, A.-R. Mohamed, Hybrid speech recognition with deep bidirectional lstm, in: IEEE Workshop on Automatic Speech Recognition and Understanding (2013) pp. 273–278.
[36] C. Gurrin, A.F. Smeaton, D. Byrne, N. Hare, G.J. Jones, N. Connor, An examination of a large visual lifelog, Information Retrieval Technology, Springer, 2008, pp. 537–542.
[37] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[38] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, K. Wood, SenseCam: A retrospective memory aid, in: ACM Conf. on Ubiquitous Computing (2006) pp. 177–193.
[39] R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, A. Kapadia, Privacy behaviors of lifeloggers using wearable cameras, in: ACM Intl. Joint Conf. on Pervasive and Ubiquitous Computing (2014) pp. 571–582.
[40] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, K. Saenko, Learning to reason: end-to-end module networks for visual question answering, in: IEEE International

Conference on Computer Vision (ICCV) (2017).
[41] IDC, IDC Forecasts Shipments of Wearable Devices to Nearly Double by 2021 as Smart Watches and New Product Categories Gain Traction. Tech. rep. International Data Corporation, 2017.
[42] Y. Iwashita, A. Takamine, R. Kurazume, M.S. Ryoo, First-person animal activity recognition from egocentric videos, in: IAPR International Conference on Pattern Recognition (ICPR) (2014).
[43] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, ACM International Conference on Multimedia (MM), ACM, 2014, pp. 675–678.
[44] V. Kalnikaite, A. Sellen, S. Whittaker, D. Kirk, Now let me see where I was: Understanding how lifelogs mediate memory, in: ACM CHI Conference on Human Factors in Computing Systems (CHI). (2010) pp. 2045–2054.
[45] S. Karim, A. Andjomshoaa, A. Tjoa, Exploiting SenseCam for helping the blind in business negotiations, Computers Helping People with Special Needs, Springer, 2006.
[46] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) pp. 3128–3137.
[47] A. Karpathy, J. Johnson, F.-F. Li, Visualizing and understanding recurrent networks. arXiv:1506.02078 (2015).
[48] A. Karpathy, A. Joulin, F.F. Li, Deep fragment embeddings for bidirectional image sentence mapping, in: Advances in Neural Information Processing Systems (NIPS) (2014) pp. 1889–1897.
[49] J. Kerr, S.J. Marshall, S. Godbole, J. Chen, A. Legge, A.R. Doherty, P. Kelly, M. Oliver, H.M. Badland, C. Foster, Using the SenseCam to improve classifications of sedentary behavior in free-living settings, Am. J. Preven. Med. 44 (3) (2013) 290–296.
[50] R. Kiros, R. Salakhutdinov, R. Zemel, Multimodal neural language models. In: Xing, E.P., Jebara, T. (Eds.), International Conference on Machine Learning. Vol. 32 of Proceedings of Machine Learning Research. PMLR, Bejing, China (2014) pp. 595–603. URL http://proceedings.mlr.press/v32/kiros14.html.
[51] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
[52] M. Korayem, R. Templeman, D. Chen, D. Crandall, A. Kapadia, Enhancing lifelogging privacy by detecting screens, in: ACM CHI Conference on Human Factors in Computing Systems (CHI) (2016).
[53] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NIPS) (2012) pp. 1097–1105.
[54] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Baby talk: understanding and generating image descriptions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011).
[55] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, Y. Choi, Collective generation of natural image descriptions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. ACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA (2012) pp. 359–368. URL http://dl.acm.org/citation.cfm?id=2390524.2390575.
[56] S. Lee, S.P.S. Prakash, M. Cogswell, V. Ranjan, D. Crandall, D. Batra, Stochastic multiple choice learning for training diverse deep ensembles, in: Advances in Neural Information Processing Systems (NIPS) (2016) pp. 2119–2127.
[57] Y. Lee, K. Grauman, Predicting important objects for egocentric video summarization, Int. J. Comput. Vision (IJCV) 114 (1) (2015).
[58] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Workshop On Text Summarization Branches Out (2004).
[59] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, European Conference on Computer Vision (ECCV), Springer, 2014, pp. 740–755.
[60] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision (IJCV) 60 (2) (2004) 91–110.
[61] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) pp. 3242–3250. URL doi: http://dx.doi.org/10.1109/CVPR.2017.345.
[62] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013).
[63] S. Mann, J. Nolan, B. Wellman, Sousveillance: inventing and using wearable computing devices for data collection in surveillance environments, Surveillance Soc. 1 (3) (2002) 331–355.
[64] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks (m-rnn). ICLR (2015).
[65] J. Mao, W. Xu, Y. Yang, J. Wang, A.L. Yuille, Explain images with multimodal recurrent neural networks. arXiv:1410.1090 (2014).
[66] L. Miller, J. Toliver, Implementing a body-worn camera program: Recommendations and lessons learned. Tech. rep. Office of Community Oriented Policing Services, 2014.
[67] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, H. Daumé, III, Midge: generating image descriptions from computer vision detections, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. EACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA (2012) pp. 747–756. URL http://dl.acm.org/citation.cfm?id=2380816.2380907.
[68] M. Moghimi, W. Wu, J. Chen, S. Godbole, S. Marshall, J. Kerr, S. Belongie, Analyzing sedentary behavior in life-logging images, in: IEEE International Conference on Image Processing (ICIP) (2014).
[69] D.H. Nguyen, G. Marcu, G.R. Hayes, K.N. Truong, J. Scott, M. Langheinrich, C. Roduner, Encountering SenseCam: personal recording technologies in everyday life, in: ACM Intl. Conf. on Ubiquitous Computing (2009) pp. 165–174.
[70] G. O'Loughlin, S.J. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, G.D. Warrington, Using a wearable camera to increase the accuracy of dietary analysis, Am. J. Prevent. Med. 44 (3) (2013) 297–301.
[71] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Annual meeting of the Association for Computational Linguistics (2002) pp. 311–318.
[72] C.C. Park, B. Kim, G. Kim, Attend to you: Personalized image captioning with context sequence memory networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017).
[73] Y. Poleg, C. Arora, S. Peleg, Temporal segmentation of egocentric videos, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014).
[74] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops (2014) pp. 512–519.
[75] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) pp. 1179–1195.
[76] M. Ryoo, T.J. Fuchs, L. Xia, J.K. Aggarwal, L. Matthies, Robot-centric activity prediction from first-person videos: What will they do to me'. In: ACM/IEEE International Conference on Human Robot Interaction (HRI) (2015) pp. 295–302.
[77] M. Ryoo, L. Matthies, First-person activity recognition: What are they doing to me? in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013) pp. 2730–2737.
[78] M. Schuster, K. Paliwal, Bidirectional recurrent neural networks, IEEE T. Signal Process. 45 (11) (1997) 2673–2681.
[79] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. arXiv:1610.02391 (2016).
[80] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014).
[81] K. Singh, K. Fatahalian, A. Efros, Krishnacam: using a longitudinal, single-person, egocentric dataset for scene understanding tasks, in: IEEE Winter Conference on Applications of Computer Vision (WACV) (2016).
[82] M. Smith, Autographer wearable camera launches tomorrow, we go hands-off. Engadget (2013).
[83] C. Szegedy, A. Toshev, D. Erhan, Deep neural networks for object detection. In: Advances in Neural Information Processing Systems (NIPS) (2013) pp. 2553–2561.
[84] T. Takeuchi, T. Narumi, K. Nishimura, T. Tanikawa, M. Hirose, Receiptlog applied to forecast of personal consumption. In: Intl. Conf. on Virtual Systems and Multimedia (2010) pp. 79–83.
[85] T. Takeuchi, K. Suwa, H. Tamura, T. Narumi, T. Tanikawa, M. Hirose, A task-management system using future prediction based on personal lifelogs and plans, in: ACM conference on Pervasive and ubiquitous computing adjunct publication (2013) pp. 235–238.
[86] R. Templeman, M. Korayem, D.J. Crandall, A. Kapadia, Placeavoider: Steering first-person cameras away from sensitive spaces. In: Network and Distributed System Security Symposium (NDSS) (2014).
[87] R. Vedantam, C. Zitnick, D. Parikh, Cider: consensus-based image description evaluation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) pp. 4566–4575.
[88] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence–video to text. arXiv:1505.00487 (2015).
[89] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 3156–3164.
[90] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. Regh, V. Singh, Gaze-enabled egocentric video summarization via constrained submodular maximization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015a).
[91] K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention. arXiv:1502.03044 (2015b).
[92] Z. Yang, Y. Yuan, Y. Wu, W.W. Cohen, R.R. Salakhutdinov, 2016. Review networks for caption generation, in: Advances in Neural Information Processing Systems (2016) pp. 2361–2369.
[93] C. Yoo, J. Shin, I. Hwang, J. Song, Facelog: capturing user's everyday face using mobile devices, in: ACM Conf. on Pervasive and Ubiquitous Computing (2013) pp. 163–166.
[94] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) pp. 4651–4659.
[95] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. arXiv:1506.06724 (2015).