Learning Visual Storylines with Skipping Recurrent Neural Networks

Gunnar A. Sigurdsson(✉), Xinlei Chen, and Abhinav Gupta

Carnegie Mellon University, Pittsburgh, USA
gunnar@cmu.edu
http://www.github.com/gsig/srnn

Abstract. What does a typical visit to Paris look like? Do people first take photos of the Louvre and then the Eiffel Tower? Can we visually model a temporal event like "Paris Vacation" using current frameworks? In this paper, we explore how we can automatically learn the temporal aspects, or storylines of visual concepts from web data. Previous attempts focus on consecutive image-to-image transitions and are unsuccessful at recovering the long-term underlying story. Our novel Skipping Recurrent Neural Network (S-RNN) model does not attempt to predict each and every data point in the sequence, like classic RNNs. Rather, S-RNN uses a framework that skips through the images in the photo stream to explore the space of all ordered subsets of the albums via an efficient sampling procedure. This approach reduces the negative impact of strong short-term correlations, and recovers the latent story more accurately. We show how our learned storylines can be used to analyze, predict, and summarize photo albums from Flickr. Our experimental results provide strong qualitative and quantitative evidence that S-RNN is significantly better than other candidate methods such as LSTMs on learning long-term correlations and recovering latent storylines. Moreover, we show how storylines can help machines better understand and summarize photo streams by inferring a brief personalized story of each individual album.

1 Introduction

In the past few years, there has been a remarkable success in learning visual concepts [1,2] and relationships [1,3] from images and text on the web. In theory, this allows the creation of systems that, given enough time and resources, can grow to know everything there is to learn. However, most of these approaches are still largely centered around single images and focus on learning static semantic relationships such as *is-part-of* [1], *is-eaten-by* [3] *etc.* Moreover, many semantic concepts have not only a visual aspect but also a temporal aspect or even storylines associated with them. For example, a visual representation of *Wedding* would involve guests entering the venue, followed by exchange of rings and finally

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46454-1_5) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 71–88, 2016.
DOI: 10.1007/978-3-319-46454-1_5

Fig. 1. Given a concept, our algorithm can automatically learn both its visual and temporal aspect (storylines) from the web. To do this, we retrieve related albums from Flickr and apply our S-RNN model to automatically discover long-term temporal patterns. Here is a visualization of the storylines the model learned for the concept *Paris*. For visualization, we distill the top images that a trained S-RNN model prefers by sampling storylines from a *Paris* photo album. Denoting the images as nodes in a graph, we visualize the most common pairwise transitions using arrowed lines. On the right, we sample three probable storylines (A,B,C) that include these 10 images. We can see that the *Eiffel Tower* is prominent early in the story followed by sightseeing of common landmarks (*Arc de Triomphe* and others) and finally visiting the *Louvre*. On a map of Paris, the *Eiffel Tower* and the *Arc de Triomphe* are indeed in close proximity

celebrations in the wedding reception. How can we learn such visual storylines from the web as well?

There are two aspects to these storylines: the visual aspect, often represented by modes in visual appearances, and the temporal aspect, which is the temporal order in which these modes appear. How do we capture both of these aspects from the web data? User photo albums in Flickr are a perfect example of web data that capture both aspects. First, most Flickr images are supplied with sufficiently informative tags, like *Paris* [4]. Second, meta-information like time is usually available. In particular, the photos in each album are taken in ordered sequences, which hypothetically embed common storylines for concepts such as *Paris*. Therefore, we propose to utilize Flickr photo streams across thousands of users and learn underlying visual storylines associated with these concepts. What is the right representation for these storylines and how do we learn it?

Recently, there has been momentous success in using CNN [5] features along with Recurrent Neural Networks [6–11] (RNNs) to represent those temporal dynamics in data [12–19]. We aim to extend that idea to modeling the dynamics in storylines. In theory, RNN can model any sequence, but has limited memory

Fig. 2. Given a concept, such as *Wedding*, our algorithm can retrieve an ordered collection of images to describe that concept (Sect. 4.3). In this figure we show the collections discovered by our model for two concepts. For example, for *Wedding* (first row), it picks images that represent four steps: guests enter; ceremony begins, marriage and celebration. For travel-related concepts like *London*, it prefers iconic landmarks for the story. The subtitles are manually provided for visualization. This is distilled from 1000 photo albums. More examples are provided in the appendix.

in practice, and can only learn short-term relationships due to vanishing gradients [20].

Our Skipping Recurrent Neural Network (S-RNN) skips through the photo sequences to extract the common latent stories, instead of trying to predict each and every item in the sequence. This effectively alleviates the artifacts of short-term correlations[1] (*e.g.* repetition) between consecutive photos in the stream, and focuses the learning effort towards the underlying story. This solution is complementary to, and different from, more complex RNN architectures such as LSTMs [21] that still focus on learning transitions between consecutive images. Similar to clustering, the S-RNN model can be efficiently trained in an *unsupervised* manner to learn a global storyline and infer a private story for each album. Different from most clustering techniques, S-RNN inherits the power of RNNs that can capture the temporal dynamics in the data.

We evaluate the effectiveness of our storyline model by comparing the storylines with baselines. In addition we evaluate the storyline model on two applications: (a) image prediction [22,23]; and (b) photo album summarization [24–27]. Constructing a convincing storyline for a concept of interest requires both visual and temporal aspects. Therefore, algorithms need to retrieve a diverse collection of images, with the right ordering among them. For *image prediction*, we show that our model is particularly suited for discovering the long-term correlations buried under the short-term repetitions in Flickr albums, while other approaches do not. Finally in the *summarization* task, the goal is to take images in a single photo album and select a small summary of those. A typical example is a series

[1] In our Flickr dataset, 71.1% of consecutive images are above average (cosine) similarity.

of photos, taken by a family on their visit to *Paris*, visiting all the iconic landmarks, such as the Eiffel Tower. Classically, summarization is approached by collecting a dataset of videos/albums and their associated summaries generated by people [26,28–31], in order to learn how to make a summary in a supervised way. This process is, however, considerably laborious. In this work, we specifically experiment with the hypothesis that a quality summary of an album can be constructed by exploiting the similarities across thousands of similar albums (*e.g. Paris*). Then a summary of the album is inferred by telling a personalized version of the story.

Contributions. (a) We present a new way of approaching sequence modeling with RNNs, by exploring all ordered subsets of the data to avoid short-term correlations between consecutive elements in the sequence. (b) We present the novel S-RNN architecture that efficiently implements this idea on web-scale datasets. (c) We demonstrate that this method can learn visual storylines for a concept (*e.g. Paris*) from the web, by showing state-of-the-art results on selecting representative images, long-term image prediction, and summarizing photo albums.

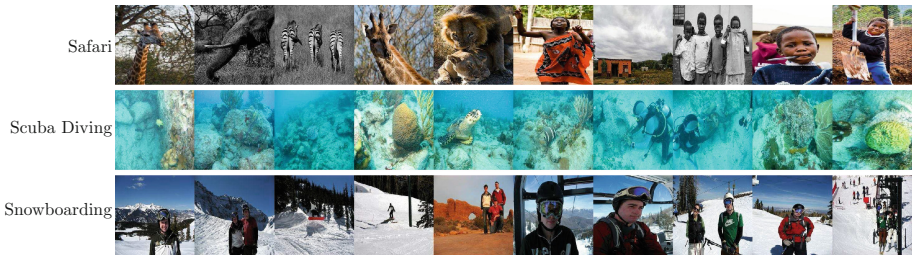

Fig. 3. Given an individual photo album, our algorithm can summarize the photo album with a ordered collection of images that capture the album in terms of its underlying concept, by first learning about the concept from thousands of albums. (Sect. 4.5). In this figure we show the summaries generated for three photo albums. One about a *Safari*, the second about *Scuba Diving*, and the third *Snowboarding*. More examples are provided in the appendix.

2 Related Work

Learning storylines. The earliest form of storyline can be traced back to the 1970-80s, where *scripts* [32] (structured representations of events, causation relationships, participants, *etc.*) are used as knowledge backbones for tasks like text summarization and question answering. Unfortunately, these rich knowledge structures require hand construction by the experts, which fundamentally limits their usage in an open domain. This motivates the recent developments of *unsupervised* approaches that can learn underlying storylines automatically [33,34] from text. Inspired by this idea, our work aims to acquire the temporal aspect of a concept automatically from images. Similar work in vision is limited by either

the scale of the data [29,35] or the domain to which the approach is applied [36]. Perhaps the most similar work is [22,23], where the storyline graphs are learned for Flickr albums. However, our work differs in several important aspects. First, while [22] is an important step in learning storylines, it focuses its learning effort on each and every pairwise transition, but our method learns the long-term latent story. In fact, [22] could be extended using this framework, but here we extend a standard RNN model. Second, our method requires no a-priori clustering, feature independence, nor a Markov assumption, and does parameters sharing like RNNs.

Temporal visual summarization. Summarizing video clips is an active area of research [37]. Many approaches have been developed seeking cues ranging from low-level motion and appearances [24,25,38] to high level concepts [26,39] and attentions [40]. This line of research has been recently extended to photo albums, and more external factors are considered for summarization besides the narrative structure. For example, in [41] the authors put forward three criteria: quality, diversity, and coverage. Later, in [42] a system is proposed that considers the social context (*e.g.* characters, aesthetics) into the summarization framework. Sadeghi *et al.* [28] also consider if a photo is memorable or iconic. Moreover, most of these approaches are *supervised*, namely the associated summaries for videos/albums are first collected by crowd-sourcing, then a model is learned to generate good summaries. While performance-wise it may seem best to leverage human supervision and external factors when available, practically it suffers serious issues like scalability and inconsistency in the ground-truth collection process, and generalizablility when applied to other domains. On the other hand, the task of summarization will be less ambiguous if the concept is given, which is exactly what we want to explore in this work.

Sequential learning with RNNs. Recurrent neural networks [6] are a subset of neural networks that can carry information across time steps. Compared to other models for sequential modeling (*e.g.* hidden Markov models, linear dynamic systems), they are better at capturing the long-range and high-order time-dependencies, and have shown superior performance on tasks like language modeling [43] and text generation [44]. In this work we extend the network to model high dimensional trajectories in videos and user albums through the space of continuous visual features. Interestingly, since our network is trained to predict images several steps away, it can be viewed as a simple and effective way to learn long term memories [21] and predict context [45] as well. Fundamentally, LSTM still looks at only the next image and decides if it should be stored it in memory, but S-RNN reasons over all future images, and decides which it should store in memory (greedy vs. global). We outperform multiple LSTM baselines in our results. Furthermore, running LSTMs directly on high-dimensional continuous features is non-trivial, and we present a network that accomplishes that.

3 Learning Visual Storylines

Given hundreds of albums for a concept, our goal is to learn the underlying visual appearances and temporal dynamics simultaneously. Once we have learned this by building upon state-of-the-art tools, we can use it for multiple storyline tasks, and distill the explicit knowledge as needed, such as in Fig. 1. In this section, we explain our novel S-RNN architecture that is trained over all ordered subsets of the data, and show that this can be accomplished with update equations equally efficient to original RNN. The full derivation of these update equations by using the EM-method is presented in the appendix. We formulate the storyline learning problem as learning an S-RNN. To understand S-RNN, we start by introducing the basic RNN model.

3.1 Recurrent Neural Networks

The basic form of RNN [6] models a time sequence by decomposing the probability of a complete sequence into sequentially predicting the next item given the history (in our application, this sequence is images in a temporal order). Given a sequence of T images $\mathbf{x}_{1:T} = \{x_1, \ldots, x_T\}$,[2] the network is trained to maximize the log-likelihood:

$$\begin{aligned} \mathcal{M}^* &= \arg\max_{\mathcal{M}} \log \mathrm{P}(\mathbf{x}_{1:T}; \mathcal{M}) - \lambda \mathcal{R}(\mathcal{M}) \\ &\text{where } \log \mathrm{P}(\mathbf{x}_{1:T}; \mathcal{M}) = \sum_t \log \mathrm{P}\left(x_{t+1} | \mathbf{x}_{1:t}; \mathcal{M}\right). \end{aligned} \tag{1}$$

Here $\mathcal{M}$ is the set of all model parameters, and $\mathcal{R}(\cdot)$ is the regularizer (*e.g.* ℓ_2). The probability $\mathrm{P}(\cdot|\cdot,\cdot)$ is task dependent, *e.g.* for language models it directly compares the soft-max output y_t with the next word x_{t+1} [43]. The standard optimization algorithm for RNNs is Back Propagation Through Time [46,47] (BPTT), a variation of gradient ascent where the gradient is aggregated through time sequences.

The model consists of three layers: input, recurrent, and output. The input layer uses the input x_t to update the hidden recurrent layer h_t using weights $\mathbf{W}_I$. The recurrent layer h_t updates itself via $\mathbf{W}_R$ and predicts the output $\mathbf{y}_t$ via weights $\mathbf{W}_O$. The update function at step t writes as follows:

$$h_t = \sigma\left(\mathbf{W}_I x_t + \mathbf{W}_R h_{t-1}\right); \quad y_t = \zeta(\mathbf{W}_O h_t). \tag{2}$$

Here $\sigma(\cdot)$ and $\zeta(\cdot)$ are non-linear activation functions, *e.g.* sigmoid, soft-max, rectified linear units [5], *etc.* All the history in RNN is stored in the memory h_t. This assumes conditional independence of x_{t+1} and $\mathbf{x}_{1:t}$ given h_t.

In practice, the recurrent layer h_t has limited capacity and the error cannot be back propagated effectively (due to vanishing gradients [20]). This can be a

[2] For simplicity in notation, we assume a single training sequence, but in our experiments we use multiple albums for one concept to discover *common* latent storylines.

critical issue for modeling sequences like photo streams—due to the high correlation between consecutive images, where the dominant pattern in the short term is *repetition*. For example, people can take multiple pictures of the same object (*e.g.* the Eiffel Tower or family members), or the entire album is about things that are visually similar (*e.g.* artwork in the Louvre or fireworks). This pattern is so salient that if an RNN is directly trained on these albums, the signals of underlying storylines are largely suppressed. How to resolve this issue of learning long-term patterns? One way is to regularize RNN with a diversity term [41]. However, note that if an album is indeed single-themed, we still want visually similar images in the storyline. Furthermore, Flickr tags are not perfect and noise in the album set can easily distract the model.

3.2 Skipping Recurrent Neural Networks

We now build upon the RNN framework to propose a skipping recurrent neural network model. Instead of learning each consecutive transition, S-RNN chooses to learn a "higher-level" version of the story, and focuses its learning effort accordingly. The key underlying idea is to select the storyline nodes by skipping a lot of images in the album and then modeling the transitions between the images selected as nodes.

Formally, let us suppose $\mathbf{x}_{1:T}$ represents the T images in the album, $\mathbf{z}_{1:N}$ is the set of indexes that represent the selected images for the storyline and the constant N is the number of nodes in the storyline. Note that $N \ll T$, $z_n \in \{1, 2, \ldots, T\}$, and $z_n < z_{n+1}$ since $\mathbf{z}$ defines an ordered subset. Our goal is to learn the maximum likelihood model parameters ($\mathcal{M}$) by maximizing the marginal likelihood of the observed data. Therefore, our objective function is:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} \log \sum_{\mathbf{z}_{1:N}} \mathrm{P}(\mathbf{x}_{1:T}, \mathbf{z}_{1:N}; \mathcal{M}) - \lambda \mathcal{R}(\mathcal{M}). \tag{3}$$

We can factorize $\mathrm{P}(\mathbf{x}_{1:T}, \mathbf{z}_{1:N}; \mathcal{M})$ as $\mathrm{P}(\mathbf{x}_{1:T}|\mathbf{z}_{1:N}; \mathcal{M})\mathrm{P}(\mathbf{z}_{1:N})$ where $\mathrm{P}(\mathbf{z}_{1:N})$ is a prior on $\mathbf{z}$. As described above, we use a simple prior on $\mathbf{z}$ that it is an ordered subset. In this work, we make an assumption that the likelihood of a whole album is proportional to the likelihood of the selected sub-sequence of images $\mathbf{x}_\mathbf{z}$ (that is, we assume $\mathrm{P}(\mathbf{x}_{1:T}|\mathbf{z}; \mathcal{M}) \propto \mathrm{P}(\mathbf{x}_\mathbf{z}; \mathcal{M})$). Factorizing, and inserting this assumption into Eq. 3 we have:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} \log \sum_{\mathbf{z}_{1:N}} \left(\prod_n \mathrm{P}\left(\mathbf{x}_{z_{n+1}} | \mathbf{x}_{\mathbf{z}_{1:n}}; \mathcal{M}\right) \right) \mathrm{P}(\mathbf{z}_{1:N}) - \lambda \mathcal{R}(\mathcal{M}). \tag{4}$$

We observe that this equation is starting to look similar to standard RNN (Eq. 1).

Maximizing the S-RNN Objective. Maximizing the marginal likelihood over all possible subsets of $\mathbf{z}$ is computationally intractable. Therefore, we make use of the Expectation Maximization (EM) algorithm, and then sequentially factor the update equations. More details of the EM derivation are given in the

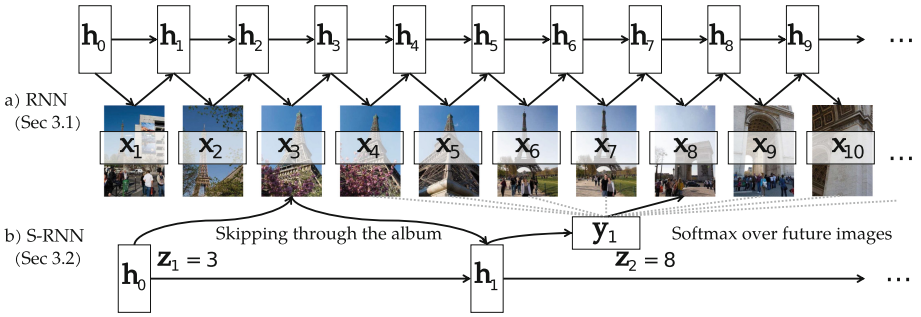

Fig. 4. Our S-RNN model (unrolled in time). Instead of trying to predicting each and every photo in the sequence (as in the basic RNN model), latent variables $\mathbf{z}_n$ are introduced into our model to skip through the photo sequences, which is an effective strategy to address the local repetition issue (multiple pictures are taken for a single object like the Eiffel Tower) and can help extract common latent stories in the entire set of albums related to a concept (*e.g. Paris*). To overcome the high-dimensional regression problem, the loss is an softmax loss over future images

appendix. During the E-step, we sample $\mathbf{z}$ given the current model, and use that to train the model in the M-step, as we would an RNN. We initialize the EM-algorithm by setting $\mathbf{z}$ based on a randomly ordered subsets of images.

S-RNN Implementation Details. Now that we know how to optimize the objective, the only design choice left is the loss $\mathrm{P}\left(\mathbf{x}_{z_{n+1}}|\mathbf{x}_{z_{1:n}};\mathcal{M}\right)$ (the data likelihood in Eq. 4). While Gaussian likelihood is often used for real-valued regression, we recognize that the space of allowed future images is not infinite, but simply images after x_{z_n}, defined as $\mathcal{X}_n$. Thus the likelihood is defined as a *softmax likelihood over the future images*:

$$\mathrm{P}\left(x_{z_{n+1}}|\mathbf{x}_{\mathbf{z}_{1:n}};\mathcal{M}\right) = \frac{\exp(\mathbf{y}_n^T x_{z_{n+1}})}{\sum_{x\in\mathcal{X}_n}\exp(\mathbf{y}_n^T x)} \tag{5}$$

where $\mathbf{y}_n$ is the output of the network after step n. Effectively, this avoids modeling the negative world as "everything except the ground truth" and instead models the negative world as "other possible choices". This significantly helps with high-dimensional data ($fc7$ features), since the possible image choices in an album are usually only few hundred, but visual features few thousand.

In summary, during training and testing, $\mathbf{z}$ is sequentially sampled using the current model (which skips through the sequence), and during training those samples used to sequentially update the network with BPTT to maximize the objective. A visualization of the idea can be found in Fig. 4. A full implementation of is available.

4 Experiments

Since there has been so little done in the area of learning storyline models and their applications, there are no established datasets, evaluation methodologies, or even much in terms of relevant previous work to compare against. Therefore, we will present our evaluation in two parts: (a) first, in Sect. 4.3, we directly evaluate how "good" our learned storyline model is. Specifically, we ask the Amazon Mechanical Turk (AMT) users how good our storyline model is compared to a baseline in terms of the representativeness and the diversity of image nodes in the storyline model; (b) next, in Sects. 4.4 and 4.5, we evaluate our storyline model for two applications: long-term prediction and album summarization. For these tasks, we show qualitative, quantitative, and user studies to demonstrate the effectiveness of S-RNN based storyline model. We begin by describing our data collection process and the baselines.

4.1 Flickr Albums Dataset

We gather collections of photo albums by querying Flickr through the YFCC100M dataset [48], a recently released public subset of the Flickr corpus containing 99.3 million images with all the meta-information like tags and time stamps. This dataset is an unrefined subset of images on Flickr, making it a reproducible way of working with web data. The selection process gathers at most 1000 photo albums for a single concept (*e.g. Paris*), with an average size of 150 images. Each album is sorted based on a photo's date taken. We experimented with seven concepts: *Christmas*, *London*, *Paris*, *Wedding*, *Safari*, *Scuba-diving*, and *Snowboarding* with a total number of 700k images. Examples from the dataset are provided in the appendix. This subset will be made available.

4.2 Implementation Details

We compare our S-RNN model with several approaches to demonstrate its effectiveness in learning visual storylines. For fairness, all the methods used the same *fc7* features from AlexNet [5] pre-trained on ImageNet.

For **S-RNN**, the *fc7* features are directly fed into the model. The network is trained with BPTT, which unrolls the network, and uses gradient ascent with a momentum of 0.9. We set the starting learning rate as 0.05, and gradually reduce it when the likelihood on the validation set no longer increases. The input size of the layer is set to 4096 (size of *fc7*), and the hidden recurrent layer size 50. We keep $N = 10$ for all the concepts as a good compromise between content and brevity (The appendix contains analysis of different sizes of N). We choose ℓ_2 regularization and set weight decay λ to be 10^{-7}. Training takes approximately 2–3 h on a single CPU. Each story was generated by sampling from the model 500 times, and picking the sampled sequence with the highest likelihood. The code is available at github.com/gsig/srnn.

Below we list the main baselines, and note that additional baselines will be added for individual experiments when necessary.

Sample. We uniformly sample from the data distribution.

K-Means. To take advantage of the global storylines shared in a concept, we apply K-Means to all the albums (similar to the first step of [22] except with different features).

Graph. We adapted the original code for [22] to use *fc7* features. Then a storyline is generated with the forward-backward algorithm as described in [22].

RNN. This architecture is similar to a language model [43] except it predicts the cluster (as in *K-Means*) of the next image. We sample without replacement to generate the story. This is a standard application of RNN to the problem.

LSTM. We train an LSTM network [49], similar to the RNN baseline.

LSTMsub. LSTM trained as before, but when generating the summary, we first generate a longer sequence ($N = 100$) and then sub-sample that sequence to the desired summary length 10. Intuitively, if LSTM was indeed able to learn the long-term correlations regardless of the repetitions, this should perform well.

S-RNN-. For ablation analysis, we also provide a baseline where we use the network without skipping, but with the softmax loss over future images. All the hyper-parameters for training are kept identical to our model except the network predicts each and every item in the sequence. This is similar to RNN, but benefits from our improved loss.

D-RNN. Similar to S-RNN-, except trained on a diverse subset of each album using the k-means++ algorithm [50]. This was significantly better than other variants, including training on random subsets, fixed interval subsets, or a random diverse subset.

4.3 Evaluating Storylines

In the first experiment, we directly evaluate how "good" the learned storyline model is for a given concept. We define the goodness of a storyline model in terms of how representative and diverse the selected images are for a given concept. Two qualitative examples for *Wedding* and *London* are shown in Fig. 2. Figure 5 shows more examples of learned storylines for different concepts. Our storyline model captures the essence of scuba-diving, snowboarding, etc., by capturing representative and diverse images (e.g., beer, fun and snowboarding during day).

Setup. For each concept, we have each method select only 10 images from 50 photo albums (thousands of photos) that best describe the concept, and AMT workers select which one they prefer. Each algorithm has access to the full training data to train the model. For Graph and RNN-based baselines, we sample multiple times from each album and use the highest ranked collection in terms of likelihood. Sample and K-Means are simply applied on all images, and in K-Means we assign the closest image to each cluster center. The appendix contains more qualitative examples.

Table 1. *Evaluating Storylines.* Fraction of the time our S-RNN storylines are preferred against competing baselines. 50 % is equal preference. Our method significantly outperforms the baselines, being preferred 60% of the time against the strongest baseline. See Sect. 4.3 for details

	K-Means	Sample	Graph	LSTM	LSTMsub	D-RNN	RNN	S-RNN-
S-RNN	71.2 %	68.3 %	79.8 %	84.3 %	70.9 %	60.0 %	85.1 %	75.5 %

Fig. 5. *Evaluating Storylines.* Images selected by S-RNN for three storylines from thousands of images for the concepts *Safari*, *Scuba Diving*, and *Snowboarding*

Results. Table 1 summarizes the results. Each comparison was given to 15 separate AMT workers. We can see that S-RNN is preferred **60 % of the time against the strongest baseline** across all the concepts. Different baselines fail in different ways. For example, Sample and K-Means can capture a diverse set of images to represent the concept, but are prone to the inherent noise in the Flickr albums. On the other hand, Graph and LSTM overfit to the short-term correlations in the data and select repetitive images. Finally, S-RNN outperforms D-RNN since S-RNN is not restricted to a single specific diversity method as in D-RNN.

4.4 Task1: Prediction

Next, we evaluate our storyline models for two applications. The first application we consider is the *prediction* task. There are two possible prediction goals: short-term prediction and long-term prediction. Short-term prediction can be considered as prediction of the next image in the album. This was the task used in [22,23]. In the case of long-term prediction, we predict the next representative event. In the case of *Paris* vacation, if the current event is Eiffel tower, the next likely event would be visiting the Trocadero. In the case of *Wedding*, if the current event is the ring ceremony, then the next representative event is the kiss of the newlyweds.

Setup. For the short-term prediction, the ground-truth is the next image in the album. But how do we collect ground-truth for long-term prediction? We ask experts to summarize the albums (hoping that album summaries will suppress short-term correlations and capture only representative events). Now we can

reformulate long-term prediction as predicting the next image in the human-generated *summary* of the album. We collected 10 ground truth summaries on average for each concept from volunteers familiar with the concepts (such as *Paris*, and *London*). Each summary consists of 10 images from a photo album that capture what the album was about. This was used as ground truth only for evaluation. Two settings are compared, the first one (labeled "*long-term*") predicts the next image in a *summary* ($N = 198$ over 10 folds each); and the second one (labeled "*short-term*") predicts the next image in the original photo album ($N = 1742$). The problem is posed as a classification task choosing from the true image, and four other images selected uniformly at random from the same album. Here we also consider **NN** that simply picks the nearest neighbor, and **FI** that picks the furthest image from the given image, both in cosine distance of *fc7* features. K-Means is not suitable for this task since it does not include temporal information. All methods were trained in an unsupervised manner for each concept as before.

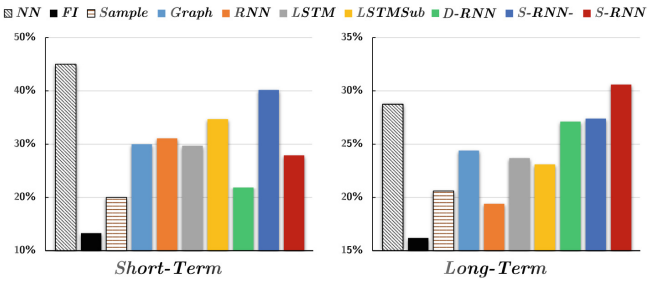

Fig. 6. *Predicting the next image.* S-RNN is best at capturing long-term correlations, and nearest-neighbors is best at capturing short-term correlations, as expected

Results. In Fig. 6 we present results for the prediction of the next image. When we consider long-term interactions between images, S-RNN successfully predicts the next image in the storyline **31** % of the time, significantly higher than baselines. On the other hand, we can see that when we simply want to predict consecutive images, NN is the best. To further visualize the results for "*long-term*" correlations, we also give example comparisons with baseline methods in Fig. 7.

4.5 Task2: Photo Album Summarization

In the final experiment, we evaluate on the task of album summarization. In particular, we focus on summarizing an individual album based on the concept (*e.g.* a *Paris* album), rather than heuristics such as image quality or presence of faces [28,41,42]. This experiment addresses the question whether storylines can help to summarize an album.

Human Generated Summaries. Photo album summarization is inherently a subjective and difficult task. To get a sense of the difficulty, we first compared

Fig. 7. *Long-term prediction.* Examples of the images predicted by our method compared to baselines. The image is chosen from a line-up of five images from the same album (generated by experts as summaries). We see our method captures *Santa→Tree* and *Closed Presents→Open Presents* while the baselines focus on similar images

the human summaries (used in Sect. 4.4) to baselines with a separate AMT preference study. We had two findings. First, for some concepts, such as *Wedding*, the albums are frequently already summaries by professional photographers, and thus generating summaries is trivial. Specifically, there is no significant difference between human generated summaries and uniformly sampling from the data distribution (Sample). We thus only evaluate on concepts where there is significant difference between human generated summaries and ones generated by baselines. Second, we found human generated summaries are only preferred **59.5** % of the time against the strongest baseline.

Setup. The photo albums for a given concept are randomly divided into a training set and a validation set with a ratio of 9:1, and no ground truth summaries were provided. We additionally consider the baseline **Local** where K-Means clustering is used for summarization by applying clustering on *fc7* features for each individual album. As before, we assign the closest image to the cluster center for clustering-based methods. While it is not required for S-RNN, we sort the selected photos in temporal order as a post-processing step for all the baselines when necessary for fair comparison.

Qualitative Results. The results for a few concepts are presented in Fig. 8. We can see that S-RNN captures a set of relevant images without losing diversity. In contrast, Local captures only diversity, and LSTM that tries to learn short-term correlations between consecutive images, and as result often prefers similar images in a row. Additional summaries by S-RNN are presented in Fig. 3.

Quantitative Evaluation. To directly compare the quality of the generated summaries, another AMT preference study was conducted. For S-RNN and each baseline, 200 random pairwise comparisons were generated. Each question was given to 5 separate workers for consistency. We used a consensus approach where a comparison gets a score of 1 if there is a tie, or a score of 2 if there is consensus.

Fig. 8. Examples of summaries generated by our method and two representative baselines for *Scuba-diving* and *Snowboarding*. In the *Scuba-diving* example Local aims to capture diversity, and thus our method is more relevant. In *Snowboarding*, LSTM focuses on short-term correlations, and chooses many similar images, while our method effectively captures the album

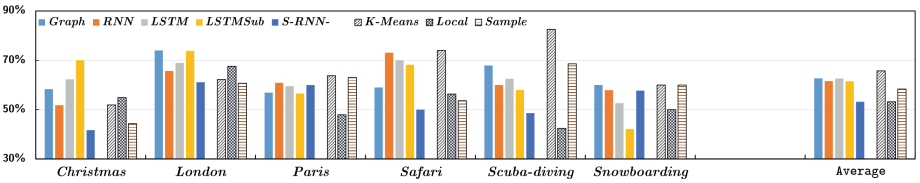

Fig. 9. Photo album summarization. AMT pairwise preference between our method and multiple baselines. 70 % means that summaries by our method were preferred 70 % of the time. It is important to keep in mind that compared to the strongest baseline, a human generated summary was on average only preferred 59.5 % of the time. Section 4.5 contains a detailed explanation of the experiment setup and analysis of the results

In Fig. 9 we present comparison with the baselines. We can see that on average our method is preferred over all the baselines. To provide a more detailed analysis, we divide the baseline methods into two groups: the *Storyline* group (filled with pure colors) that captures the latent temporal information in the data, and the Non-Storyline group (filled with patterns) that do not. The *Storyline* group includes Graph, RNN, LSTM, LSTMSub, and S-RNN- (Our method also falls into this group), while the *Non-Storyline* group has K-Means, Local and Sample. There are few interesting points:

1. S-RNN performs relatively better on travel-related albums (*Paris*, *London*) suggesting it is easier to latch onto landmarks than high-level concepts like in *Christmas*.
2. For concepts like *Christmas*, methods that learn short-term correlations from the data distribution are still preferred by the users. The fact that S-RNN- outperforms LSTMs and RNNs, can be interpreted as follows. RNNs suffer from the curse of dimensionality if naively applied to storyline learning, but the S-RNN loss reduces the dimensionality of the output space by an order of magnitude (4096 to 100s).

3. While simple as they seem, Local and Sample are very competitive baselines. We believe the reason is that Local aims to provide a diverse set of images from each album, and Sample is representative of the underlying data. Therefore, with the post-processing step that re-arranges the selected images in temporal order, these methods can do well on good albums. However, they do poorly when the album is noisy, as illustrated in Fig. 8 first example.

Does Time Information Help Summarization? For further analysis, we compared the described S-RNN with S-RNN trained on shuffled data (ordering discarded) with a preference study on AMT. S-RNN using the time information was preferred 68.4 % over S-RNN without time information, demonstrating that the time information significantly helps to generate a summary liked by people.

Transferring storyline knowledge. Each album can have different stories and themes. In Fig. 10 we present two different summaries of two photo albums. The first album is a *Scuba Diving* album, and the first summary from that album is generated with the model trained on *Scuba Diving* albums. In the second row, the same album is summarized using a model trained on *Wedding* albums. We can see that this emphasizes scenic beach pictures reminiscent of a beach resort wedding. The second album is a *Paris* album, and the first summary is generated using *Paris* model. The second summary however, is generated using a *Christmas* model, and we can see that this emphasizes pictures of churches and sparkling lights at night.

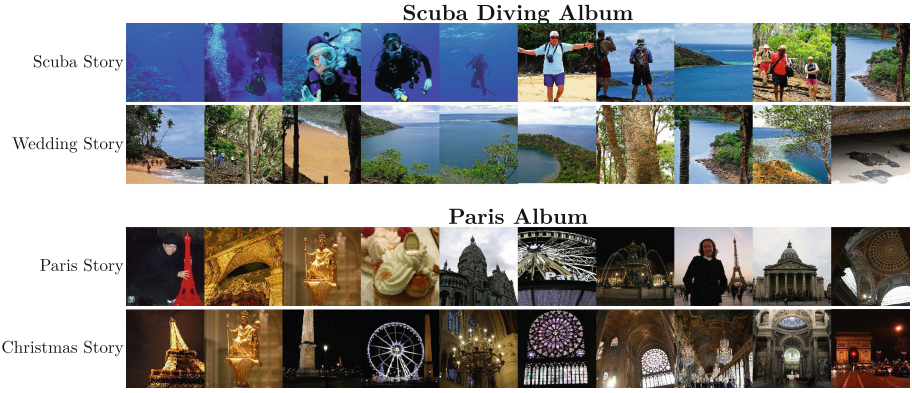

Fig. 10. The first two rows show a *Scuba Diving* album summarized with a *Scuba* model and a *Wedding* model, and the last two show a *Paris* album summarized with a *Paris* model and a *Christmas* model. The *Wedding* story emphasizes the beach resort images of the *Scuba* album, and the *Christmas* story emphasizes the churches and sparkling lights images in the *Paris* album

5 Conclusion

We have presented an approach to learn visual storylines for concepts automatically from the web. Specifically, we use Flickr albums and train an S-RNN

model to capture the long-term temporal dynamics for a concept of interest. The model is designed to overcome the challenges posed by high correlations between consecutive photos in the album if sequence predictors are directly applied. We evaluate our model on learning storylines, image prediction and album summarization, and show both qualitatively and quantitatively that our method excels at both extracting salient visual signals for the concept, and learning long-term storylines to capture the temporal dynamics.

Acknowledgements. This research was supported by the Yahoo-CMU InMind program, ONR MURI N000014-16-1-2007, and a hardware grant from Nvidia. The authors would like to thank Olga Russakovsky and Christoph Dann for invaluable suggestions and advice, and all the anonymous reviewers for helpful advice on improving the manuscript.

References

1. Chen, X., Shrivastava, A., Gupta, A.: NEIL: extracting visual knowledge from web data. In: ICCV (2013)
2. Divvala, S.K., Farhadi, A., Guestrin, C.: Learning everything about anything: webly-supervised visual concept learning. In: CVPR (2014)
3. Sadeghi, F., Divvala, S.K., Farhadi, A.: VisKE: visual knowledge extraction and question answering by visual verification of relation phrases. In: CVPR (2015)
4. Izadinia, H., Farhadi, A., Hertzmann, A., Hoffman, M.D.: Image classification and retrieval from user-supplied tags (2014). arXiv preprint: arXiv:1411.6909
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
6. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
7. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: a neural-based approach to answering questions about images. In: ICCV, pp. 1–9 (2015)
8. Shih, K.J., Singh, S., Hoiem, D.: Where to look: focus regions for visual question answering (2015). arXiv preprint: arXiv:1511.07394
9. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering (2015). arXiv preprint: arXiv:1511.02274
10. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: grounded question answering in images (2015). arXiv preprint: arXiv:1511.03416
11. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering (2016). arXiv preprint: arXiv:1603.01417
12. Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions (2014). arXiv preprint: arXiv:1412.2306
13. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)
14. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR (2015)
15. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R.J., Darrell, T., Saenko, K.: Sequence to sequence - video to text (2015). arXiv preprint: arXiv:1505.00487
16. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention (2015). arXiv preprint: arXiv:1502.03044

17. Gregor, K., Danihelka, I., Graves, A., Wierstra, D.: Draw: a recurrent neural network for image generation (2015). arXiv preprint: arXiv:1502.04623
18. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books (2015). arXiv preprint: arXiv:1506.06724
19. Chen, X., Zitnick, C.L.: Learning a recurrent visual representation for image caption generation. In: CVPR (2015)
20. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. TNN **5**(2), 157–166 (1994)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
22. Kim, G., Xing, E.P.: Reconstructing storyline graphs for image recommendation from web community photos. In: CVPR (2014)
23. Kim, G., Sigal, L., Xing, E.P.: Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In: CVPR (2014)
24. DeMenthon, D., Kobla, V., Doermann, D.: Video summarization by curve simplification. In: ACM MM, pp. 211–218. ACM (1998)
25. Ngo, C.W., Ma, Y.F., Zhang, H.J.: Video summarization and scene detection by graph modeling. TCSVT **15**(2), 296–305 (2005)
26. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: CVPR (2013)
27. Martin-Brualla, R., He, Y., Russell, B.C., Seitz, S.M.: The 3D jigsaw puzzle: mapping large indoor spaces. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part III. LNCS, vol. 8691, pp. 1–16. Springer, Heidelberg (2014)
28. Sadeghi, F., Tena, J.R., Farhadi Ali, S.L.: Learning to select and order vacation photographs. In: WACV (2015)
29. Xiong, B., Kim, G., Sigal, L.: Storyline representation of egocentric videos with an applications to story-based search. In: ICCV, pp. 4525–4533 (2015)
30. Kim, G., Moon, S., Sigal, L.: Joint photo stream and blog post summarization and exploration. In: CVPR, pp. 3081–3089. IEEE (2015)
31. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: video summarization by visual co-occurrence. In: CVPR, pp. 3584–3592 (2015)
32. Shank, R., Abelson, R.: Scripts, plans, goals and understanding (1977)
33. Chambers, N., Jurafsky, D.: Unsupervised learning of narrative event chains. In: ACL (2008)
34. McIntyre, N., Lapata, M.: Learning to tell tales: a data-driven approach to story generation. In: ACL (2009)
35. Wang, D., Li, T., Ogihara, M.: Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In: AAAI (2012)
36. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: CVPR (2009)
37. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. TOMCCAP **3**(1), 3 (2007)
38. Cernekova, Z., Pitas, I., Nikou, C.: Information theory-based shot cut/fade detection and video summarization. TCSVT **16**(1), 82–91 (2006)
39. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR (2012)
40. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: ACM MM (2002)

41. Sinha, P., Mehrotra, S., Jain, R.: Summarization of personal photologs using multidimensional content and context. In: ICMR (2011)
42. Obrador, P., De Oliveira, R., Oliver, N.: Supporting personal photo storytelling for social albums. In: ACM MM, pp. 561–570. ACM (2010)
43. Mikolov, T.: Recurrent neural network based language model. In: INTERSPEECH (2010)
44. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: ICML (2011)
45. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
46. Williams, R.J., Zipser, D.: Gradient-based learning algorithms for recurrent networks and their computational complexity. In: Back-Propagation: Theory, Architectures and Applications, pp. 433–486 (1995)
47. Werbos, P.J.: Generalization of backpropagation with application to a recurrent gas market model. Neural Netw. **1**(4), 339–356 (1988)
48. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: The new data and new challenges in multimedia research (2015). arXiv preprint: arXiv:1503.01817
49. Karpathy, A., Johnson, J., Li, F.: Visualizing and understanding recurrent networks (2015). arXiv preprint: arXiv:1506.02078
50. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, pp. 1027–1035 (2007)