Learning Prototypical Event Structure from Photo Albums

Antoine Bosselut[†], **Jianfu Chen**[‡], **David Warren**[‡], **Hannaneh Hajishirzi**[†] and **Yejin Choi**[†]
[†]Computer Science & Engineering, University of Washington, Seattle, WA
{antoineb, hannaneh, yejin}@cs.washington.edu
[‡]Department of Computer Science, Stony Brook University, Stony Brook, NY
{jianchen, warren}@cs.stonybrook.edu

Abstract

Activities and events in our lives are structural, be it a vacation, a camping trip, or a wedding. While individual details vary, there are characteristic patterns that are specific to each of these scenarios. For example, a wedding typically consists of a sequence of events such as walking down the aisle, exchanging vows, and dancing. In this paper, we present a data-driven approach to learning event knowledge from a large collection of photo albums. We formulate the task as constrained optimization to induce the prototypical temporal structure of an event, integrating both visual and textual cues. Comprehensive evaluation demonstrates that it is possible to learn multimodal knowledge of event structure from noisy web content.

1 Introduction

Many common scenarios in our lives, such as a wedding or a camping trip, show characteristic structural patterns. As illustrated in Figure 1, these patterns can be sequential, such as in a wedding, where exchanging vows generally happens before cutting the cake. In other scenarios, there may be a set of composing events, but no prominent temporal relations. A camping trip, for example, might include events such as hiking, which can happen either before or after setting up a tent.

This observation on the prototypical patterns in everyday scenarios goes back to early artificial intelligence research. Scripts (Schank and Abelson, 1975), an early formalism, were developed to encode the necessary background knowledge to support an inference engine for common sense reasoning in limited domains. However, early approaches based on hand-coded symbolic representations proved to be brittle and difficult to scale.

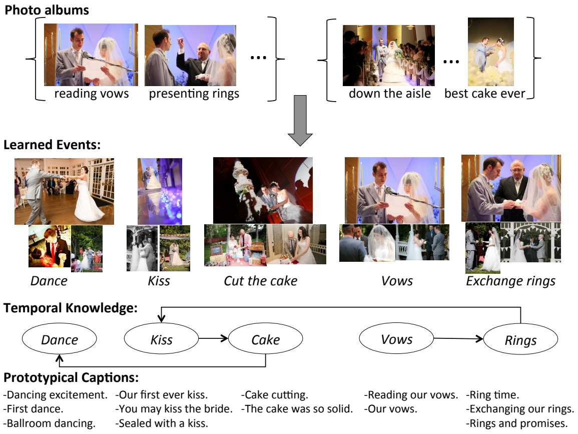

Figure 1: We collect photo albums of common scenarios (e.g., weddings) and cluster their images and captions to learn the hierarchical events that make up these scenarios. We use constrained optimization to decode the temporal order of these events, and we extract the prototypical descriptions that define them.

An alternative direction in recent years has been statistical knowledge induction, i.e., learning script or common sense knowledge bottom-up from large-scale data. While most prior work is based on text (Pichotta and Mooney, 2014; Jans et al., 2012; Chambers and Jurafsky, 2008; Chambers, 2013), recent work begins exploring the use of images as well (Bagherinezhad et al., 2016; Vedantam et al., 2015).

In this paper, we present the first study for learning knowledge about common life scenarios (e.g., weddings, camping trips) from a large collection of online photo albums with time-stamped images and their captions. The resulting dataset includes 34,818 time-stamped photo albums corresponding to 12 distinct event scenarios with 1.5 million images and captions (see Table 1 for more details).

We cast unsupervised learning of event structure as a sequential multimodal clustering problem, which requires solving two subproblems concurrently: identifying the boundaries of events and assigning identities to each of these events. We formulate this process as constrained optimization, where constraints encode the temporal event patterns that are induced directly from the data. The outcome is a statistically induced prototypical structure of events characterized by their visual and textual representations.

We evaluate the quality and utility of the learned knowledge in three tasks: temporal event ordering, segmentation prediction, and multimodal summarization. Our experimental results show the performance of our model in predicting the order of photos in albums, partitioning photo albums into event sequences, and summarizing albums.

2 Overview

The high-level goal of this work is unsupervised induction of the prototypical event structure of common scenarios from multimodal data. We assume a two-level structure: high-level events, which we refer to as *scenarios* (e.g., wedding, funeral), are given, and low-level events (e.g., dance, kiss, vows), which we refer to as *events*, are to be automatically induced. In this section, we provide the overview of the paper (Section 2.1), and introduce our new dataset (Section 2.2).

2.1 Approach

Given a large collection of photo albums corresponding to a scenario, we want to learn three aspects of event knowledge by (1) identifying events common to the given scenario (Section 4.1), (2) learning temporal relations across events (Section 4.2), and (3) extracting prototypical captions for each event (Section 4.3).

To induce the prototypical event structure, an important subproblem we consider is individual photo album analysis, where the task is (1) partitioning each photo album into a sequence of segments, and (2) assigning the event identity to each segment. We present an inference model based on Integer Linear Programming (ILP) in Section 3 to perform both segmentation and event identification simultaneously, in consideration of the learned knowledge that we describe in Section 4.

Finally, we evaluate the utility of the automatically induced knowledge in the context of three concrete tasks: temporal ordering of photos (Section 6.1), album segmentation (Section 6.2), and photo album summarization (Section 6.3).

scenario	# of albums	# of images
Wedding	4689	192K
Marathon	3961	158K
Cooking	1168	36K
Funeral	781	28K
Barbecue	735	22K
Baby Birth	688	21K
Paris Trip	4603	306K
New York Trip	4205	267K
Camping	4063	159K
Thanksgiving	5928	153K
Christmas	3449	98K
Independence Day	548	22K
Total	34,818	1.5M

Table 1: Dataset Statistics: the number of albums and images compiled for each scenario. The middle horizontal line separates the scenarios we predict have a well-defined temporal structure (top) from those we predict do not (bottom).

2.2 Dataset

For this study, we have compiled a new corpus of multimodal photo albums across 12 distinct scenarios. It comprises of 34,818 albums containing 1.5 million pairs of online photographs and their textual descriptions. Table 1 shows the list of scenarios and the corresponding data statistics. We choose six scenarios (the top half of Table 1) that we expect have an inherent temporal event structure that can be learned and six that we expect do not (the bottom half of Table 1).

The dataset is collected using the Flickr API[1,2]. We use the scenario names and variations of them (e.g., Paris Trip, Paris Vacation) as queries for images. We then form albums from these images by grouping images by user, sorting them by timestamp, and extracting groups that are within a contained time frame (e.g., 24 hours for a wedding, 5 days for a trip). For all images, we extract the first sentences of the corresponding textual descriptions as captions and also store their timestamps. This data is publicly available at `https://www.cs.washington.edu/projects/nlp/protoevents`.

3 Inference Model for Multimodal Event Segmentation and Identification

Given a photo album, the goal of the inference is to assign events to photos and to segment albums by event. More formally, given a sequence of M photos $P = \{p_1, \ldots, p_M\}$, and N learned events $E = \{e_1, \ldots, e_N\}$, the task is to assign each photo to a

[1] https://www.flickr.com/services/api/

[2] https://pypi.python.org/pypi/flickrapi/1.4.5

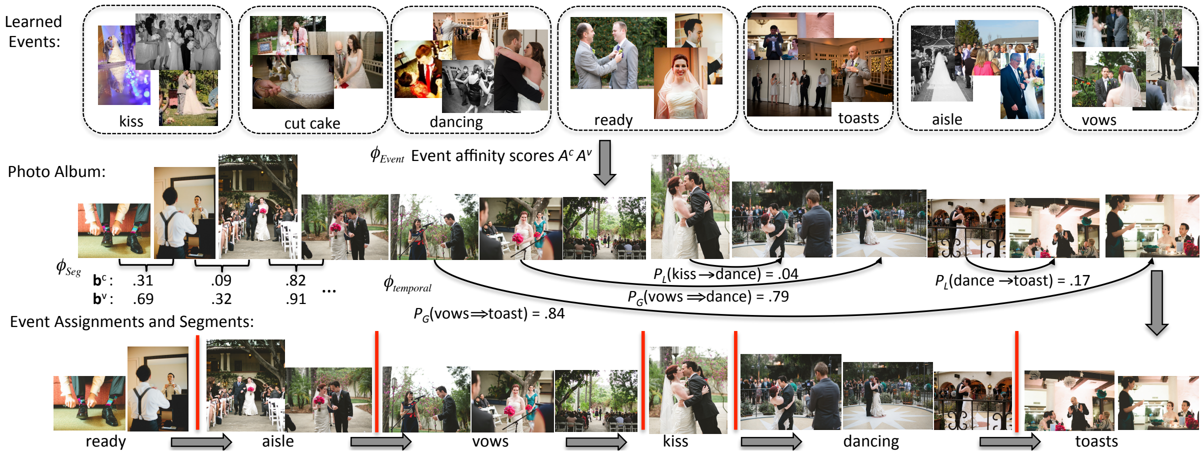
Learned Events:
kiss
cut cake
dancing
ready
toasts
aisle
vows
ϕ_{Event} Event affinity scores $A^c A^v$
Photo Album:
ϕ_{Seg}
$\mathbf{b}^c$: .31 .09 .82 ...
$\mathbf{b}^v$: .69 .32 .91
$\phi_{temporal}$
P_L(kiss → dance) = .04
P_G(vows ⇒ dance) = .79
P_L(dance → toast) = .17
P_G(vows ⇒ toast) = .84
Event Assignments and Segments:
ready
aisle
vows
kiss
dancing
toasts

Figure 2: The events learned in Section 4 are assigned to photos based on textual (A^c) and visual (A^v) affinities, which encode how well a photo represents an event (ϕ_{event}). Segmentation scores (ϕ_{seg}) between adjacent photos encourage similar photos to be assigned the same event. Local transition, P_L, and global pairwise ordering, P_G, probabilities encode the learned temporal knowledge between events. $\phi_{temporal}$ encourages event assignments toward a learned temporal structure of the scenario.

single event. The event assignment can be viewed as a latent variable for each photo. We formulate a constrained optimization (depicted in Figure 2) that maximizes the objective function, F, which consists of three scoring components: (a) event assignment scores ϕ_{event} (Section 3.1), (b) segmentation scores ϕ_{seg} (Section 3.2), and (c) temporal knowledge scores $\phi_{temporal}$ (Section 3.3):

$$F = \phi_{event} + \phi_{seg} + \phi_{temporal} \tag{1}$$

Decision Variables. The binary decision variable $\mathbf{X}_{i,k}$ indicates that photo p_i is assigned to event e_k. The binary decision variable $\mathbf{Z}_{i,j,k,l}$ indicates that photos p_i and p_j are assigned to events e_k and e_l, respectively:

$$\mathbf{Z}_{i,j,k,l} := \mathbf{X}_{i,k} \wedge \mathbf{X}_{j,l} \tag{2}$$

3.1 Event Assignment Scores

Event assignment scores quantify the textual and visual affinity between a photo p_i and an event e_k. Affinities are measures of representation similarity between photos and events. These scores push photos displaying a certain event to be assigned to that event. For now we assume the textual affinity matrix $\mathbf{A}^c \in [0,1]^{M\times N}$ and the visual affinity matrix $\mathbf{A}^v \in [0,1]^{M\times N}$ are given. We describe how we obtain these affinity matrices in Section 4.1. Event assignment scores are defined as the weighted sum of both textual and visual affinity:

$$\phi_{event} = \sum_{i=1}^{M}\sum_{k=1}^{N}\Big(\gamma_{ce}\mathbf{A}^c_{i,k} + \gamma_{ve}\mathbf{A}^v_{i,k}\Big)\mathbf{X}_{i,k} \tag{3}$$

where $\mathbf{X}_{i,k}$ is a photo-event assignment decision variable, and γ_{ce} and γ_{ve} are hyperparameters that balance the contribution of both affinities.

3.2 Segmentation Scores

Segmentation scores quantify textual and visual similarities between adjacent photos. These scores encourage similar adjacent photos to be assigned to the same event. We define a similarity score between adjacent photos equal to the weighted sum of their textual ($\mathbf{b}^c$) and visual ($\mathbf{b}^v$) similarities:

$$\phi_{seg} = \sum_{i=1}^{M-1}\sum_{k=1}^{N}\Big(\gamma_{cs}\mathbf{b}^c_i + \gamma_{vs}\mathbf{b}^v_i\Big)\mathbf{Z}_{i,i+1,k,k} \tag{4}$$

where $\mathbf{b}^c$, $\mathbf{b}^v \in [0,1]^{(M-1)\times 1}$ are vectors of textual and visual similarity scores between adjacent photos whose i^{th} element corresponds to the similarity score between photos p_i and p_{i+1}, $\mathbf{Z}$ is a decision variable defined by Equation 2, and γ_{cs} and γ_{vs} are hyperparameters balancing the contribution of both types of similarity. The similarity scores in the $\mathbf{b}$ vectors are computed using cosine similarity of the feature representations of adjacent images in both the textual and visual modes.

3.3 Temporal Knowledge Scores

Temporal knowledge scores quantify the compatibilities across different event assignments in terms of their relative ordering. For now, we assume two types of temporal knowledge matrices are given: $\mathbf{L} \in [0,1]^{N\times N}$ which stores local transition probabilities for every pair of events, e_k and e_l, and $\mathbf{G} \in [0,1]^{N\times N}$ which stores global pairwise ordering probabilities for every pair of events, e_k

and e_l. We describe how we obtain these temporal knowledge matrices in Section 4.2. The temporal knowledge score, defined below, encourages the inference model to assign events that are compatible with the learned temporal knowledge:

$$\begin{aligned} \phi_{temporal} &= \gamma_{lp} \sum_{i=0}^{M} \sum_{k,l=1}^{N} \mathbf{L}_{k,l} \mathbf{Z}_{i,i+1,k,l} \\ &+ \gamma_{gp} \sum_{i=1}^{M} \sum_{j=i}^{M} \sum_{k,l=1}^{N} \mathbf{G}_{k,l} \mathbf{Z}_{i,j,k,l} \end{aligned} \tag{5}$$

where $\mathbf{Z}$ is a decision variable defined by Equation 2, and γ_{lp} and γ_{gp} are hyperparameters that balance the contribution of local and global temporal knowledge in the objective.

3.4 Constraints

We include hard constraints that force each photo to be assigned to exactly one event:

$$\sum_{k=1}^{N} \mathbf{X}_{i,k} = 1 \tag{6}$$

The number of these constraints is linear in the number of photos in an album. We also include hard constraints to ensure consistencies among binary decision variables $\mathbf{X}$ and $\mathbf{Z}$:

$$\frac{1}{2}(\mathbf{X}_{i,k} + \mathbf{X}_{j,l}) - \mathbf{Z}_{i,j,k,l} \geq 0 \tag{7}$$

which states that $\mathbf{Z}_{i,j,k,l}$ can be 1 only if both $\mathbf{X}_{i,k}$ and $\mathbf{X}_{j,l}$ are 1. The number of constraints for segmentation scores and local transition probabilities is $O(MN^2)$ because they model interactions between adjacent photos for all event pairs. The number of these constraints for global pairwise ordering probabilities is $O(M^2N^2)$ because they model interactions between all pairs of photos in an album for all event pairs.

4 Learned Event Knowledge

We learn base events for each scenario by clustering photos from training albums related to that scenario (Figure 3). As described in Section 3, these base events and their temporal knowledge are incorporated in a joint model for event induction in unseen albums.

4.1 Learned Event Representation

We perform k-means clustering over captions to create a base event model. We perform text-only clustering at first since visual cues are significantly noisier. Because not all photos have informative captions, it is expected that this base clustering will form meaningful clusters only over a subset of the data. For each scenario, the largest cluster corresponds to the "miscellaneous" cluster as the captions in it tend to be relatively uninformative about specific events. This cluster is excluded when computing temporal knowledge probabilities (Section 4.2).

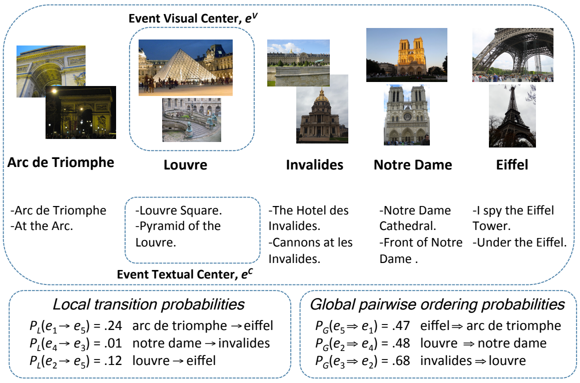
Event Visual Center, e^V
Arc de Triomphe
Louvre
Invalides
Notre Dame
Eiffel
-Arc de Triomphe
-At the Arc.
-Louvre Square.
-Pyramid of the Louvre.
-The Hotel des Invalides.
-Cannons at les Invalides.
-Notre Dame Cathedral.
-Front of Notre Dame .
-I spy the Eiffel Tower.
-Under the Eiffel.
Event Textual Center, e^C
Local transition probabilities
$P_L(e_1 \rightarrow e_5) = .24$ arc de triomphe → eiffel
$P_L(e_4 \rightarrow e_3) = .01$ notre dame → invalides
$P_L(e_2 \rightarrow e_5) = .12$ louvre → eiffel
Global pairwise ordering probabilities
$P_G(e_5 \Rightarrow e_1) = .47$ eiffel ⇒ arc de triomphe
$P_G(e_2 \Rightarrow e_4) = .48$ louvre ⇒ notre dame
$P_G(e_3 \Rightarrow e_2) = .68$ invalides ⇒ louvre

Figure 3: Photos are clustered by their captions. We can compute the visual, e_k^v, and caption, e_k^c, centers for all the clusters, as well as the local transition, P_L, and global pairwise ordering, P_G, probabilities between these events based on the sequential patterns they exhibit in the training set.

The visual and textual representations of an event are computed using the average of the visual and textual features, respectively, of photos assigned to that event. We compute each textual affinity $\mathbf{A}_{i,k}^c$ in the event assignment scores (Equation 3) as the cosine similarity between the textual features of the caption for photo p_i and the textual representation of event e_k. For textual features, we extract noun and verb unigrams using TurboTagger (Martins et al., 2013) and weigh them by their discriminativeness relative to their scenario, $P(S|w)$. Given scenario S and word w, $P(S|w)$ is defined as the number of albums for the scenario the word occurs in divided by the total number of albums in that scenario. The visual affinity $\mathbf{A}_{i,k}^v$ is the similarity between the visual features of photo p_i and the visual representation of event e_k. For visual features, we use the convolutional features from the final layer activations of the 16-layer VGGNet model (Simonyan and Zisserman, 2015).

4.2 Temporal Knowledge

Local transition probabilities. These probabilities, denoted as P_L, encode an expected sequence of events using temporal patterns among adjacent

Wedding		Camping		Funeral	
aisle	Walking down the aisle Bride walking down the aisle	*tent*	Inside our tent Setting up the tent	*service*	Graveside service The service
vow	Exchanging vows Reading the vows Reciting vows to each other	*fire*	Building the Fire Getting the Fire going Around the Fire	*pay respect*	Paying Respects Respect
dance	First Dance Everybody Dancing Dancing the Night Away	*sunset*	Watching the Sunset Sunset from camp Sunset on the first night	*goodbye*	Saying Goodbye

Table 2: Sample learned events and prototypical captions

photos. We model P_L for each pair of events as a multinomial distribution,

$$P_L(e_k \to e_l) = \frac{C(e_k \to e_l)}{\sum_{m=1}^{N} C(e_k \to e_m)} \quad (8)$$

where C is the observed counts of that specific event transition. This is the likelihood that an event e_k is immediately followed by event e_l.

Global pairwise ordering probabilities. These probabilities, denoted as P_G, encode global structural patterns about events. We model P_G for each pair of events as a binomial distribution by computing the likelihood that an event occurs before another at any point in an album,

$$P_G(e_k \Rightarrow e_l) = \frac{C(e_k \Rightarrow e_l)}{C(e_k \Rightarrow e_l) + C(e_l \Rightarrow e_k)} \quad (9)$$

where $C(e_k \Rightarrow e_l)$ is the observed counts of e_k occurring anytime before e_l in all photo albums. These global probabilities model relations among events assigned to all photos in the album, not just events assigned to photos that are adjacent to one another. This distinction is important because these probabilities can encode global patterns between events and are not limited to modeling a sequential event chain.

We use these learned temporal probabilities, P_L and P_G, in matrices $\mathbf{L}$ and $\mathbf{G}$ from $\phi_{temporal}$ (Equation 5). These matrices are used to index local transition probabilities and global pairwise ordering probabilities for pairs of events when computing temporal knowledge scores in the inference model (Section 3.3).

4.3 Prototypical Captions

After clustering the photos, the representative language of the captions in each cluster begins to tell a story about each scenario. The event names are automatically extracted using the most common content words among captions in the cluster. For each cluster, we also compile *prototypical captions* by extracting captions whose lemmatized forms are frequently observed throughout multiple albums in the scenario. Sample events and their prototypical captions from three scenarios are displayed in Table 2.

5 Experimental Setup

Data split. For scenarios with more than 1000 albums, we use 100 albums for each of the development and test sets and use the rest for training. For scenarios with less than 1000 albums, we use 50 albums for each of the development and test sets, and the rest for training.

Implementation details. We optimize our objective function using integer linear programming (Roth and Yih, 2004) with the Gurobi solver (Inc., 2015). For computational efficiency, temporally close sets of consecutive photos are treated as one unit during the optimization. We use these units to reduce the number of variables and constraints in the model from a function of the number of photos to a function of the number of units. We form these units heuristically by merging images agglomeratively when their timestamps are within a certain range of the closest image in a unit. When merging photos, the textual affinity of each unit for a particular event is the maximum affinity for that event among photos in that unit. The visual affinity of each unit is the average of all affinities for that event among photos in that unit. The textual and visual similarities of consecutive units are defined in terms of the similarities between the two photos at the units' boundary. Temporal information for events not aligned well with a particular unit should not influence the objective, so we include temporal scores only for unit-event pairs which have both textual and visual event assignment scores greater than 0.05.

Hyperparameters. We tune the hyperparameters using grid search on the development set. In models where the corresponding objective components are included, we set $\gamma_{ce} = 1$, $\gamma_{ve} = 1$, $\gamma_{cs} = .5$, $\gamma_{vs} = .15$, $\gamma_{lp} = 1$, and $\gamma_{gp} = \frac{4}{Q}$ (where Q is the

Model	Wedding	Baby Birth	Marathon	Cooking	Funeral	Barbecue	Indep. Day	Camping	Thanksgiving	Paris Trip	NY Trip	Christmas
k-MEANS	52.7	52.5	53.8	53.0	50.5	50.2	53.2	52.3	51.4	**53.1**	**50.3**	51.4
NO TEMPORAL	58.6	66.3	62.6	56.5	50.8	51.7	**58.0**	52.6	54.3	51.7	49.1	50.9
FULL MODEL	**60.0**	**66.5**	**64.5**	**63.2**	**53.1**	**58.6**	56.0	**55.5**	**56.1**	52.3	48.5	**52.4**

Table 3: Temporal ordering pairwise photo results. The metric reported is accuracy, the percentage of time the correct photo is picked as coming first based on the event assigned to it. Scenarios with an expected temporal structure are in the left half of the table.

number of event units). For k-means clustering, we use 10 random restarts and 40 cluster centers for the WEDDING, CAMPING, PARIS TRIP, and NY TRIP scenarios. For all other scenarios, we use 30 cluster centers.

6 Experimental Results

We evaluate the performance of our model on three tasks. The first task evaluates the effect of learned temporal knowledge in predicting the correct order of photos in an unseen album (Section 6.1). The second task evaluates the model's ability to segment albums into logical groupings (Section 6.2). The third task evaluates the quality of prototypical captions and their use in photo album summarization (Section 6.3).

6.1 Temporal Ordering of Photos

We evaluate the model's ability to capture the temporal relationships between events in the scenario. Given two randomly selected photos p_i and p_j from an album, the task is to predict which of the photos appears earlier in the album using their event assignments. We compare the full model that assigns events to photos using ILP (Section 3) with two baselines: k-MEANS , which assigns events to photos using k-means clustering over captions (Section 4), and NO TEMPORAL: a variant of the full model that does not use temporal knowledge scores ($\phi_{temporal}$ in Equation 1) for optimization.

We run each method over a test photo album, in which the events e_k and e_l are assigned to the photos p_i and p_j, respectively. We then use the learned global pairwise ordering probabilities (Section 4.2) to predict which photo appears earlier in the album. We report the accuracy of each method in predicting the order of photos compared to the actual order of photos in the albums. We perform this experiment 50 times for each album and average the number of correct choices across every album and every trial.

Results. Table 3 reports the results of the full model compared to the baselines. The results show that temporal knowledge generally helps in predicting photo ordering. We observe that the full model achieves higher scores for scenarios for which we expect would have a sequential structure (e.g., WEDDING, BABY BIRTH, MARATHON). Conversely, the full model achieves lower overall scores in non-sequential scenarios (e.g., PARIS TRIP, NEW YORK TRIP). Qualitatively, we notice interesting temporal patterns such as the fact that during a marathon, the starting line occurs before the medal awards with 92.3% probability, or that Parisian tourists have a 24% chance ($\sim 10\times$ higher than random chance) of visiting the Eiffel Tower immediately after the Arc de Triomphe (a high local transition probability that correctly implies their real world proximity).

6.2 Album Segmentation

Our model partitions photos in albums into coherent events. The album segmentation evaluation tests if the model recovers the same sequences of photos that a human would identify in a photo album as events.

Evaluation. We had an impartial annotator label where they thought events began and ended in 10 candidate albums of greater than 100 photos for three scenarios: WEDDING, FUNERAL, CAMPING. We evaluate how well our model can replicate these boundaries with two metrics. The first metric is the F_1 score of recovering the same boundaries annotated by humans. The second metric is d, the difference between the number of events segmented by the model compared to the annotated albums. We report results for exact event boundaries as well as relaxed boundaries where the start of an event can be r photos away from the start of an annotated event, where r is the relaxation coefficient. For reference, we note that albums in the wedding scenario were dual annotated and the agreement between annotators is 56.9% for $r = 0$ and 77.5% for $r = 2$.

Results. Table 4 shows comparison of the the full ILP model with same baselines we described before, k-MEANS and NO TEMPORAL. The table shows that the full model generally outperforms the k-MEANS baseline for all three scenarios.

In the WEDDING scenario, the F_1 score for the full

Model	r	WEDDING		FUNERAL		CAMPING	
		F_1	d	F_1	d	F_1	d
k-MEANS	0	27.1	32.9	27.9	29.6	**31.2**	46.0
NO TEMPORAL		32.0	-5.6	**35.9**	**.9**	22.0	-15.6
FULL MODEL		**37.8**	**1.3**	32.2	4.2	27.5	**-10.1**
k-MEANS	2	40.8	32.9	38.3	29.6	46.2	46.0
NO TEMPORAL		49.6	-5.6	**57.6**	**.9**	35.4	-15.8
FULL MODEL		**57.5**	**1.3**	51.6	5.0	**51.4**	**-10.1**

Table 4: Segmentation results for our full model. F_1 scores how often our model recovers the same boundaries annotated by humans. d is the average difference between the number of events identified by the model in an album and marked by annotators. r is the relaxation coefficient.

Feature Group Excluded	P	R	F_1	d
FULL MODEL	36.7	**42.8**	**37.8**	1.3
- Visual Event Affinity	37.7	37.0	35.3	-1.7
- Textual Segmentation	37.1	41.5	37.4	.8
- Visual Segmentation	35.1	42.1	36.5	1.7
- Local Ordering Probs.	36.9	40.3	36.9	**.2**
- Global Ordering Probs.	**40.5**	25.0	29.5	-5.8

Table 5: Ablation study of objective function components for the wedding scenario. P, R, and F_1 are the precision, recall and F-measure of recovering the same boundaries annotated by humans. d is the average difference between the number of events identified by our models and the annotators.

model is consistently higher. The k-MEANS baseline oversamples the number of events in albums, which is indicated by an average d significantly greater than 0. For the FUNERAL scenario, the NO TEMPORAL baseline outperforms the full model. We attribute this difference to the smaller data subset (see Table 1) making it harder to learn the temporal relations in the scenario, which makes the contributions of the local and global temporal probabilities unexpected. In the CAMPING scenario, the F_1 score for the k-MEANS baseline is higher than that of the full model when $r = 0$. At a high-level, CAMPING is a scenario we expect has less of a known structure compared to other scenarios and may be harder to segment into its events.

Ablation Study. Table 5 depicts the performance of ablations of the full model for the wedding scenario. Results show that removing any component of the objective functions yields lower recall and F_1 scores than the full model for $r = 0$. The exception is removing local ordering probabilities, which yields a higher d. These observations support the hypothesis that all of the components of the objective function contribute to segmenting the album into subsequences of photos depicting the same event. Particularly, we note the degradation when removing the global ordering probabilities, indicating that approaches which model only local event transitions such as hidden Markov models would not be suitable for this task.

6.3 Photo Album Summarization

The final experiment evaluates how our learned prototypical captions can improve downstream tasks such as summarization and captioning.

6.3.1 Summaries

The goal of a good summary is to select the most salient pictures of an album. In our setting, a good summary should have a high coverage of the events in an album and choose the photos that most appropriately depict these events. Given a photo budget b, we choose a subset of photos that aims for these goals. To summarize a test album, we run our model over the entire album. This will yield h unique events assigned to the photos in the album. For each of these h events, we choose the photo with the highest event assignment score for that event (Equation 3) to be in the summary. If $h > b$, we count the number of photos in the training set assigned to each of the h events and choose the photos corresponding to the b events with the largest membership of photos in the training set. If $h < b$, we complete the summary with $b - h$ photos from the "miscellaneous" event that are spaced evenly throughout the album. Finally, we replace the caption of each selected photo with a prototypical caption (Section 4.3) for the assigned event.

Baseline. We evaluate against two baselines. The first baseline, KTH, involves including a photo in the summary every $k = M/b$ photos. The second baseline, k-MEANS, uses the events assigned to photos from k-means clustering and then picks b photos in the same manner as our main model.

Evaluation. We evaluate the summaries produced by each method with a human evaluation using Amazon Mechanical Turk (AMT). We use albums from the test set that contain more than 40 photos for the wedding scenario. For each album, at random, we present two summaries generated by two algorithms. AMT workers are instructed to choose the better summary considering both the images and the captions. For each comparison of two summaries for an album, we aggregate answers from three workers by majority voting. We set $b = 7$. The number of assigned events in an album, h, varies by album.

Results. As seen in Table 6, the summary from the full model is preferred 57.7% of the time compared to the KTH baseline. The summaries generated using the full model perform slightly better than the summaries from k-MEANS. We attribute

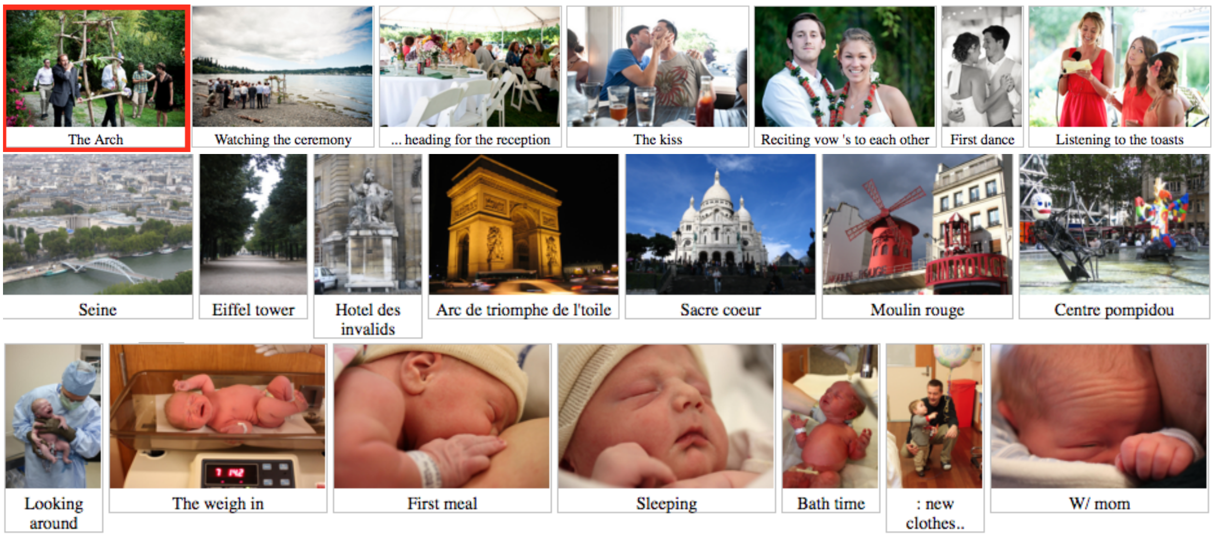

Figure 4: Example summaries from the wedding, Paris trip, and baby birth scenarios. In cases where the album had less events than b, the additionally chosen photos are outlined in red. These photos do not have their caption replaced by a prototypical captions and merely fill out the summary.

Method	Selection Rates	
Full Model vs. KTH	**57.7**	42.3
Full Model vs. k-means	**53.8**	47.2
k-means vs. KTH	**53.8**	47.2

Table 6: Summarization results. The selection rates indicate the percentage of time the corresponding method in the left-most column was picked.

Method	Scenario Relevance	Image Relevance	Grammar
LSTM	**4.90**	2.85	3.74
Full Model	4.55	**3.66**	**4.08**
Raw Captions	4.10	4.36	4.28

Table 7: Captioning results. We evaluate the caption quality of the prototypical captions of the full model, those generated by an LSTM trained on the raw captions, and original captions. Captions were evaluated on 3 metrics: grammatical correctness, how relevant they were to the scenario, and how relevant they were to their assigned image.

the superior performance of the full model to the fact that it redistributes photos with noisy captions throughout the events, allowing for a larger sample to estimate visual representations of events, yielding more accurate visual affinity measurements to choose the summarization photos. As can be seen from qualitative examples in Figure 4, the photos chosen and the captions assigned cover key events that would occur during the scenario and describe them in a coherent way. Additional examples are available at `https://www.cs.washington.edu/projects/nlp/protoevents`.

6.3.2 Prototypical Captions

We also evaluate the quality of the prototypical captions assigned to every photo in the summaries. For each album, we use the same sets of b photos from the full model in the summarization task and evaluate the quality of the prototypical captions paired with that group of photos.

Evaluation. We evaluate the quality of captions assigned to every photo by asking AMT workers to rate the captions on three different metrics: grammaticality, relevance to the scenario to which the image belongs, and relevance to its paired image. Five AMT workers rate each group of b photos on a five point Likert scale for each metric. We compare the prototypical captions for every photo in the summary with captions generated by an LSTM model[3] trained on every photo-caption pair in the training set for a scenario. We also compare with the original raw captions for each image in the summary. Because we chose photos with the highest event assignment scores (Equation 3) to be in the summary, the raw captions for this evaluation are cleaner and more descriptive than most captions in the dataset.

Results. Our model outperforms the LSTM-generated captions in the image relevance and grammaticality scores, but did worse in scenario

[3] We use a single-layer encoder-decoder LSTM. The cell state and the input embedding dimensions are 256. Visual inputs are the final layer convolutional features of the VGG-16 model and are fine-tuned during training. We use RMSprop to train the network with a base learning rate of .0001 and 30% dropout. We train the model for 45 epochs on a single NVIDIA Titan X GPU with mini batch size 100. To decode, we use beam search with beam size 5.

relevance. We attribute this result to LSTM captions having little caption variation because the model learns frequency statistics without any knowledge of latent events. Almost all LSTM captions mention the words *bride, wedding*, or *groom*, yielding a very high scenario score for the caption, even if that caption is grammatically incorrect or irrelevant to the image. As expected the raw captions have high relevance to the original image, and they are grammatical, but can be less relevant to the corresponding scenario.

7 Related Work

Previous studies have explored unsupervised induction of salient content structure in newswire texts (Barzilay and Lee, 2004), temporal graph representations (Bramsen et al., 2006), and storyline extraction and event summarization (Xu et al., 2013). Another line of research finds the common event structure from children's stories (McIntyre and Lapata, 2009), where the learned plot structure is used to stochastically generate new stories (Goyal et al., 2010; Goyal et al., 2013). Our work similarly aims to learn the typical temporal patterns and compositional elements that define common scenarios, but with multimodal integration.

Compared to studies that learn narrative schemas from natural language (Pichotta and Mooney, 2014; Jans et al., 2012; Chambers and Jurafsky, 2009; Chambers, 2013; Cassidy et al., 2014), or compile script knowledge from crowdsourcing (Regneri et al., 2010), our work explores a new source of knowledge that allows grounded event learning with temporal dimensions, resulting in a new dataset of scenario types that are not naturally accessible from newswire or literature.

While recent studies have explored videos and photo streams as a source of discovering complex events and learning their sequential patterns (Kim and Xing, 2014; Kim and Xing, 2013; Tang et al., 2012; Tschiatschek et al., 2014), their focus was mostly on the visual modality. Zhang et al. (2015) explored multimodal information extraction focusing specifically on identifying video clips that referred to the same event in television news. This contrasts to the goal of our study that aims to learn the temporal structure by which common scenarios unfold.

Integrating language and vision has attracted increasing attention in recent years across diverse tasks such as image captioning (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Fang et al., 2015; Xu et al., 2015; Chen et al., 2015), cross modal semantic modeling (Lazaridou et al., 2015), information extraction (Morency et al., 2011; Rosas et al., 2013; Zhang et al., 2015; Izadinia et al., 2015), common-sense knowledge (Vedantam et al., 2015; Bagherinezhad et al., 2016), and visual storytelling (Huang et al., 2016). Our work is similar to both common sense knowledge learning and visual story completion. Our model learns commonsense knowledge on the hierarchical and temporal event structure from scenario-specific multimodal photo albums, which can be viewed as visual stories about common life events.

Recent work focused on photo album summarization using visual (Sadeghi et al., 2015) and multimodal representations (Sinha et al., 2011). Our work identifies the nature of common events in scenarios and learns their timelines and characteristic forms.

8 Conclusion

We introduce a novel exploration to learn script-like knowledge from photo albums. We model stochastic event structure to learn both the event representations (textual and visual) and the temporal relations among those events. Our event induction method incorporates learned knowledge about events, partitions photo albums into segments, and assigns events to those segments. We show the significance of our model in learning and using learned knowledge for photo ordering, album segmentation, and summarization. Finally, we provide a dataset depicting 12 scenarios with $\sim$1.5 M images for future research. Future directions could include exploring nuances in the type of temporal knowledge that can be learned across different scenarios.

Acknowledgements

We thank the anonymous reviewers for many insightful comments and members of UW NLP for feedback and support. The work is in part supported by NSF grants IIS-1408287, IIS-1524371, IIS-1616112, DARPA under the CwC program through the ARO (W911NF-15-1-0543), the Allen Institute for AI (66-9175), the Allen Distinguished Investigator Award, and gifts by Google and Facebook.

References

Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are elephants bigger than butterflies? reasoning about sizes of objects. In *Proceedings of the Conference in Artificial Intelligence (AAAI)*.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *NAACL-HLT*.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 189–198. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *ACL*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 602–610. Association for Computational Linguistics.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Jianfu Chen, Polina Kuznetsova, David S Warren, and Yejin Choi. 2015. Déja image-captions: A corpus of expressive descriptions in repetition. In *NAACL-HLT*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86. Association for Computational Linguistics.

Amit Goyal, Ellen Riloff, et al. 2013. A computational model for plot units. *Computational Intelligence*, 29(3):466–488.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *NAACL*.

Gurobi Optimization Inc. 2015. Gurobi optimizer reference manual.

Hamid Izadinia, Fereshteh Sadeghi, Santosh K Divvala, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2015. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10–18.

Bram Jans, Steven Bethard, Ivan Vuli, and Marie-Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Gunhee Kim and Eric P Xing. 2013. Jointly aligning and segmenting multiple web photo streams for the inference of collective photo storylines. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 620–627. IEEE.

Gunhee Kim and Eric Xing. 2014. Reconstructing storyline graphs for image recommendation from web community photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3882–3889.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado, May–June. Association for Computational Linguistics.

André FT Martins, Miguel B Almeida, and Noah A Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL*.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 217–225. Association for Computational Linguistics.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM.

Karl Pichotta and Raymond J Mooney. 2014. Statistical script learning with multi-argument events. In *EACL*, volume 14, pages 220–229.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988. Association for Computational Linguistics.

Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, (3):38–45.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. Technical report, DTIC Document.

Fereshteh Sadeghi, J Rafael Tena, Ali Farhadi, and Leonid Sigal. 2015. Learning to select and order vacation photographs. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 510–517. IEEE.

Roger C Schank and Robert P Abelson. 1975. *Scripts, plans, and knowledge*. Yale University.

K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. 2011. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 4. ACM.

Kevin Tang, Li Fei-Fei, and Daphne Koller. 2012. Learning latent temporal structure for complex event detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1250–1257. IEEE.

Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. 2014. Learning mixtures of submodular functions for image collection summarization. In *Advances in Neural Information Processing Systems*, pages 1413–1421.

R. Vedantam, X. Lin, T. Batra, C. L. Zitnick, and D. Parikh. 2015. Learning common sense through visual abstraction. In *Proceedings of the International Conference in Computer Vision (ICCV)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Shize Xu, Shanshan Wang, and Yan Zhang. 2013. Summarizing complex events: a cross-modal solution of storylines extraction and reconstruction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1281–1291.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2048–2057.

Tongtao Zhang, Hongzhi Li, Heng Ji, and Shih-Fu Chang. 2015. Cross-document event coreference resolution based on cross-media features. In *Proc. Conference on Empirical Methods in Natural Language Processing*.