Sort Story: Sorting Jumbled Images and Captions into Stories

Harsh Agrawal[*,1] **Arjun Chandrasekaran**[*,1,2] **Dhruv Batra**[1] **Devi Parikh**[1] **Mohit Bansal**[2]
[1]Virginia Tech [2]TTI-Chicago
{harsh92, carjun, dbatra, parikh}@vt.edu, mbansal@ttic.edu

Abstract

Temporal common sense has applications in AI tasks such as QA, multi-document summarization, and human-AI communication. We propose the task of *sequencing* – given a jumbled set of aligned image-caption pairs that belong to a story, the task is to sort them such that the output sequence forms a coherent story. We present multiple approaches, via unary (position) and pairwise (order) predictions, and their ensemble-based combinations, achieving strong results on this task. As features, we use both text-based and image-based features, which depict complementary improvements. Using qualitative examples, we demonstrate that our models have learnt interesting aspects of temporal common sense.

1 Introduction

Sequencing is a task for children that is aimed at improving understanding of the temporal occurrence of a sequence of events. The task is, given a jumbled set of images (and maybe captions) that belong to a single story, sort them into the correct order so that they form a coherent story. Our motivation in this work is to enable AI systems to better understand and predict the temporal nature of events in the world. To this end, we train machine learning models to perform the task of "sequencing".

Temporal reasoning has a number of applications such as multi-document summarization of multiple sources of, say, news information where the relative order of events can be useful to accurately merge information in a temporally consistent manner. In question answering tasks (Richardson et al., 2013; Fader et al., 2014; Weston et al., 2015; Ren et al., 2015), answering questions related to when an event occurs, or what events occurred prior to a particular event require temporal reasoning. A good temporal model of events in everyday life, i.e., a "temporal common sense", could also improve the quality of communication between AI systems and humans.

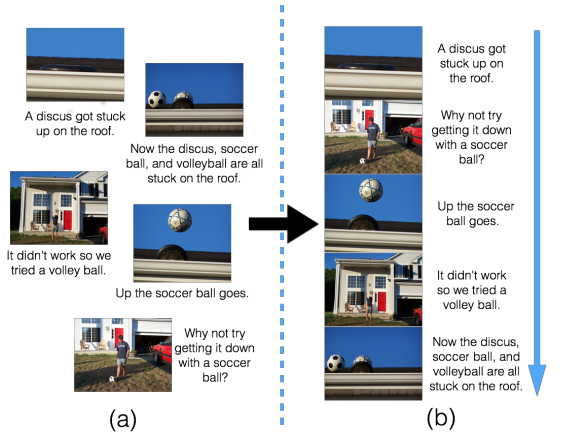


Figure 1: (a) The input is a jumbled set of aligned image-caption pairs. (b) The output is a sequence of image-caption pairs that forms a coherent story.

Stories are a form of narrative sequences that have an inherent temporal common sense structure. We propose the use of visual stories depicting personal events to learn temporal common sense. We use stories from the Sequential Image Narrative Dataset (SIND) (Ferraro et al., 2016) in which a set of 5 aligned image-caption pairs together form a coherent story. Given an input story that is jumbled (Fig. 1 (a)), we train machine learning models to sort them into a coherent story (Fig. 1(b)).[1]

[*]Denotes equal contribution

[1]Note that 'jumbled' here refers to the loss of temporal ordering; image-caption pairs are still aligned.

Our contributions are as follows:
– We propose the task of visual story sequencing.
– We implement two approaches to solve the task: one based on individual story elements to predict position, and the other based on pairwise story elements to predict relative order of story elements. We also combine these approaches in a voting scheme that outperforms the individual methods.
– As features, we represent a story element as both text-based features from the caption and image-based features, and show that they provide complementary improvements. For text-based features, we use both sentence context and relative order based distributed representations.
– We show qualitative examples of our models learning temporal common sense.

2 Related Work

Temporal ordering has a rich history in NLP research. Scripts (Schank and Abelson, 2013), and more recently, narrative chains (Chambers and Jurafsky, 2008) contain information about the participants and causal relationships between events that enable the understanding of stories. A number of works (Mani and Schiffman, 2005; Mani et al., 2006; Boguraev and Ando, 2005) learn temporal relations and properties of news events from the dense, expert-annotated TimeBank corpus (Pustejovsky et al., 2003). In our work, however, we use multimodal story data that has no temporal annotations.

A number of works also reason about temporal ordering by using manually defined linguistic cues (Webber, 1988; Passonneau, 1988; Lapata and Lascarides, 2006; Hitzeman et al., 1995; Kehler, 2000). Our approach uses neural networks to avoid feature design for learning temporal ordering.

Recently, Mostafazadeh et al. (2016) presented the "ROCStories" dataset containing 5 sentence stories with stereotypical causal and temporal relations between events. In our work though, we make use of a multi-modal story-dataset that contains *both* images and associated story-like captions.

Some works in vision (Pickup et al., 2014; Basha et al., 2012) also temporally order images; typically by finding correspondences between multiple images of the same scene using geometry-based approaches. Similarly, Choi et al. (2016) attempt to compose a story out of multiple short video clips. They define metrics based on scene dynamics and coherence, and use dense optical flow and patch-matching. In contrast, our work deals with stories containing potentially visually dissimilar but *semantically* coherent set of images and captions.

A few other recent works (Kim et al., 2015; Kim et al., 2014; Kim and Xing, 2014; Sigurdsson et al., 2016; Bosselut et al., 2016; Wang et al., 2016) summarize hundreds of individual streams of information (images, text, videos) from the web that deal with a single concept or event, to learn a common theme or *storyline* or for *timeline summarization*. Our task, however, is to predict the correct sorting of a given story, which is different from summarization or retrieval.

Tang et al. (2012) attempt to reason about the temporal structure of complex events in sport and multimedia event videos. While their motivation is similar to ours, their work deals with temporal patterns in *video* sequences. In our work, however, we attempt to learn temporal common sense from *stories*, consisting of a sequence of aligned image-caption pairs.

3 Approach

In this section, we first describe the two components in our approach: unary scores that do not use context, and pairwise scores that encode relative orderings of elements. Next, we describe how we combine these scores through a voting scheme.

3.1 Unary Models

Let $\sigma \in \Sigma_n$ denote a permutation of n elements (image-caption pairs). We use σ_i to denote the position of element i in the permutation σ. A unary score $S_u(\sigma)$ captures the appropriateness of each story element i in position σ_i:

$$S_u(\sigma) = \sum_{i=1}^{n} P(\sigma_i|i) \quad (1)$$

where $P(\sigma_i|i)$ denotes the probability of the element i being present in position σ_i, which is the output from an n-way softmax layer in a deep neural network. We experiment with 2 networks – (1) A language-alone unary model (Skip-Thought+MLP) uses a Long Short Term Memory

(LSTM) (Hochreiter and Schmidhuber, 1997) RNN encoder to embed a caption into a vector space. We use the Skip-Thought (Kiros et al., 2015) LSTM, which is trained on the BookCorpus (Zhu et al., 2015) to predict the context (preceding and following sentences) of a given sentence. These embeddings are fed as input into a Multi-Layer Perceptron (MLP).

(2) A language+vision unary model (Skip-Thought+CNN+MLP) that embeds the caption as above and embeds the image via a Convolutional Neural Network (CNN). We use the activations from the penultimate layer of the 19-layer VGG-net (Simonyan and Zisserman, 2014), which have been shown to generalize well. Both embeddings are concatenated and fed as input to an MLP.

In both cases, the best ordering of the story elements (optimal permutation) $\sigma^* = \arg\max_{\sigma \in \Sigma_n} S_u(\sigma)$ can be found efficiently in $O(n^3)$ time with the Hungarian algorithm (Munkres, 1957). Since these unary scores are not influenced by other elements in the story, they capture the semantics and linguistic structures associated with specific positions of stories *e.g.*, the beginning, the middle, and the end.

3.2 Pairwise Models

Similar to learning to rank approaches (Hang, 2011), we develop pairwise scoring models that given a pair of elements (i, j), learn to assign a score: $S([\![\sigma_i < \sigma_j]\!] \mid i, j)$ indicating whether element i should be placed before element j in the permutation σ. Here, $[\![\cdot]\!]$ indicates the Iverson bracket (which is 1 if the input argument is true and 0 otherwise). We develop and experiment with the following 3 pairwise models:

(1) A language-alone pairwise model (Skip-Thought+MLP) that takes as input a pair of Skip-Thought embeddings and trains an MLP (with hinge-loss) that outputs $S([\![\sigma_i < \sigma_j]\!] \mid i, j)$, the score for placing i before j.

(2) A language+vision pairwise model (Skip-Thought+CNN+MLP) that concatenates the Skip-Thought and CNN embeddings for i and j and trains a similar MLP as above.

(3) A language-alone neural position embedding (NPE) model. Instead of using frozen Skip-Thought embeddings, we learn a task-aware ordered distributed embedding for sentences. Specifically, each sentence in the story is embedded $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^d_+$, via an LSTM with ReLU non-linearities. Inspired by the order embedding work of Vendrov et al. (2015), we use an asymmetric penalty that encourages sentences appearing early in the story to be placed closer to the origin than sentences appearing later in the story.

$$\begin{aligned} L_{ij} &= \Big\| \max(0, \alpha - (\mathbf{x}_j - \mathbf{x}_i)) \Big\|^2 \\ Loss &= \sum_{1<=i<j=n} L_{ij} \end{aligned} \tag{2}$$

At train time, the parameters of the LSTM are learned end-to-end to minimize this asymmetric ordered loss (as measured over the gold-standard sequences). At test time, we use $S([\![\sigma_i < \sigma_j]\!] \mid i, j) = L_{ij}$. Thus, as we move away from the origin in the embedding space, we traverse through the sentences in a story. Each of these three pairwise approaches assigns a score $S(\sigma_i, \sigma_j | i, j)$ to an ordered pair of elements (i,j), which is used to construct a pairwise scoring model:

$$S_p(\sigma) = \sum_{1<=i<j<=n} \Big\{ S([\![\sigma_i < \sigma_j]\!]) - S([\![\sigma_j < \sigma_i]\!]) \Big\}, \tag{3}$$

by summing over the scores for all possible ordered pairs in the permutation. This pairwise score captures local contextual information in stories. Finding the best permutation $\sigma^* = \arg\max_{\sigma \in \Sigma_n} S_p(\sigma)$ under this pairwise model is NP-hard so approximations will be required. In our experiments, we study short sequences ($n = 5$), where the space of permutations is easily enumerable ($5! = 120$). For longer sequences, we can utilize integer programming methods or well-studied spectral relaxations for this problem.

3.3 Voting-based Ensemble

To combine the complementary information captured by the unary (S_u) and pairwise models (S_p), we use a voting-based ensemble. For each method in the ensemble, we find the top three permutations. Each of these permutations (σ^k) then vote for a particular element to be placed at a particular position. Let V be a vote matrix such that V_{ij} stores the number of votes for i^{th} element to occur at j^{th} position, *i.e.* $V_{ij} = \sum_k [\![\sigma^k_i == j]\!]$).

We use the Hungarian algorithm to find the optimal permutation that maximizes the votes assigned, *i.e.* $\sigma^*_{\text{vote}} = \arg\max_{\sigma \in \Sigma_n} \sum_{i=1}^{n} \sum_{j=1}^{n} V_{ij} \cdot [\![\sigma_i == j]\!]$. We experimented with a number of model voting combinations and found the combination of pairwise Skip-Thought+CNN+MLP and neural position embeddings to work best (based on a validation set).

4 Experiments

4.1 Data

We train and evaluate our model on personal multimodal stories from the SIND (Sequential Image Narrative Dataset) (Ferraro et al., 2016), where each story is a sequence of 5 images and corresponding story-like captions. The narrative captions in this dataset, e.g., "friends having a good time" (as opposed to "people sitting next to each other") capture a sequential, conversational language, which is characteristic of stories. We use 40,155 stories for training, 4990 for validation and 5055 stories for testing.

4.2 Metrics

We evaluate the performance of our model at correctly ordering a jumbled set of story elements using the following 3 metrics: **Spearman's rank correlation** (Sp.) (Spearman, 1904) measures if the ranking of story elements in the predicted and ground truth orders are monotonically related (higher is better). **Pairwise accuracy** (Pairw.) measures the fraction of pairs of elements whose predicted relative ordering is the same as the ground truth order (higher is better). **Average Distance** (Dist.) measures the average change in position of all elements in the predicted story from their respective positions in the ground truth story (lower is better).

4.3 Results

Pairwise Models vs Unary Models As shown in Table 1, the pairwise models based on Skip-Thought features outperform the unary models in our task. However, the Pairwise Order Model performs worse than the unary Skip-Thought model, suggesting that the Skip-Thought features which encode context of a sentence play a crucial role in predicting the ordering of story sentences. It is interesting to note that for pairwise Skip-Thought model, the avg. error in position of all elements in the predicted story is < 1.

Method	Features	Sp.	Pairw.	Dist.
Random Order		0.000	0.500	1.601
Unary	SkipThought	0.508	0.718	1.373
	SkipThought + CNN	0.507	0.718	1.375
Pairwise	SkipThought	0.546	0.732	0.923
	SkipThought + CNN	0.562	0.740	0.899
Pairwise Order	NPE	0.480	0.704	1.010
Voting	SkipThought + CNN (Pairwise) + NPE	**0.602**	**0.762**	**0.829**

Table 1: Performance of different approaches and features at the sequencing task.

Contribution of Image Features We augment both our models with image features, and observe that our pairwise model has a visible improvement in performance, unlike the unary model. The reason for this could be that a pair of images may contain signals of the relative ordering (e.g., falling after tripping), but individual images may not have affinities to specific positions in a story. This is unlike individual sentences, which may contain linguistic and semantic patterns revealing their position (beginning, middle, or end) in stories.

Ensemble Voting To exploit the fact that unary and pairwise models, as well as text and image features capture different aspects of the story, we combine them using our voting ensemble. Based on the validation set, we found that combining predictions from Pairwise Order model and Pairwise model that uses both Skip-Thought and CNN features performs the best. This voting based method achieves the best performance on all three metrics. This shows that our different approaches indeed capture complementary information regarding feasible orderings of caption-image pairs to form a coherent story.

4.4 Qualitative Analysis

Visualizations of position predictions from our model demonstrate that it has learnt the *three act structure* (Trottier, 1998) in stories – the setup, the middle and the climax. Our model has also learnt aspects of *temporal common sense*, which we demonstrate via discriminative, position-based word-cloud

figures. E.g., our model believes that a story should begin by mentioning a 'party', 'wedding', 'location' (i.e., by describing the occasion/event). It tends to associate middle sentences with words like 'people', 'friend', 'everyone', etc., and concludes with words like 'finally', 'afterwards', 'great time', 'night', etc. We also present success and failure examples of our sorting model's predictions. See appendix for more details and figures.

5 Conclusion

We propose the task of "sequencing" in a set of image-caption pairs, with the motivation of learning temporal common sense. We implement multiple neural network models based on individual and pairwise element-based predictions (and their ensemble), and utilize both image and text features, to achieve strong performance on the task. Our best system, on average, predicts the ordering of sentences to within a distance error of 0.8 (out of 5) positions. We also analyze our predictions and show qualitative examples that demonstrate temporal common sense.

6 Acknowledgements

We thank Ramakrishna Vedantam for helpful suggestions and discussions. This work was supported in part by the following: National Science Foundation CAREER awards to DB and DP, Army Research Office YIP awards to DB and DP, ICTAS Junior Faculty awards to DB and DP, Army Research Lab grant W911NF-15-2-0080 to DP and DB, Office of Naval Research grant N00014-14-1-0679 to DB, Paul G. Allen Family Foundation Allen Distinguished Investigator award to DP, Google Faculty Research award to DP and DB, AWS in Education Research grant to DB, NVIDIA GPU donations to DB and MB, an IBM Faculty Award and Bloomberg Data Science Research Grant to MB.

Appendix

A Confusion Matrix for Predicting Position of an Element

Fig. 2 shows the 5-way classification confusion matrix for our best performing method i.e., Voting ensemble of Pairwise Skip-Thought+CNN and Pairwise Order (Neural Position Embedding (NPE)). The block-diagonal matrix structure shows that the best model does a reasonable job at predicting the first and the last element of a story but is often confused by elements in the middle part of the story. Thus, at a minimum, the model has learnt the *three act structure* in stories, i.e., the setup, the middle and the climax.

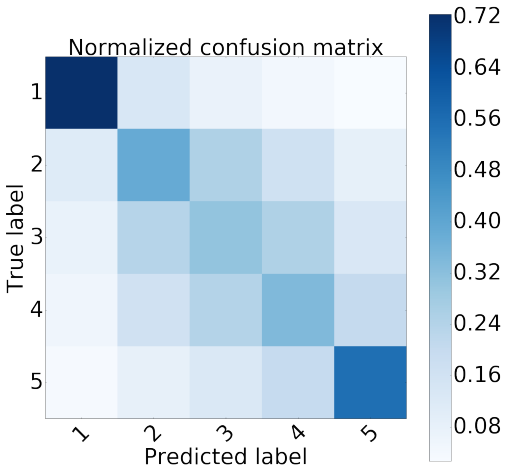


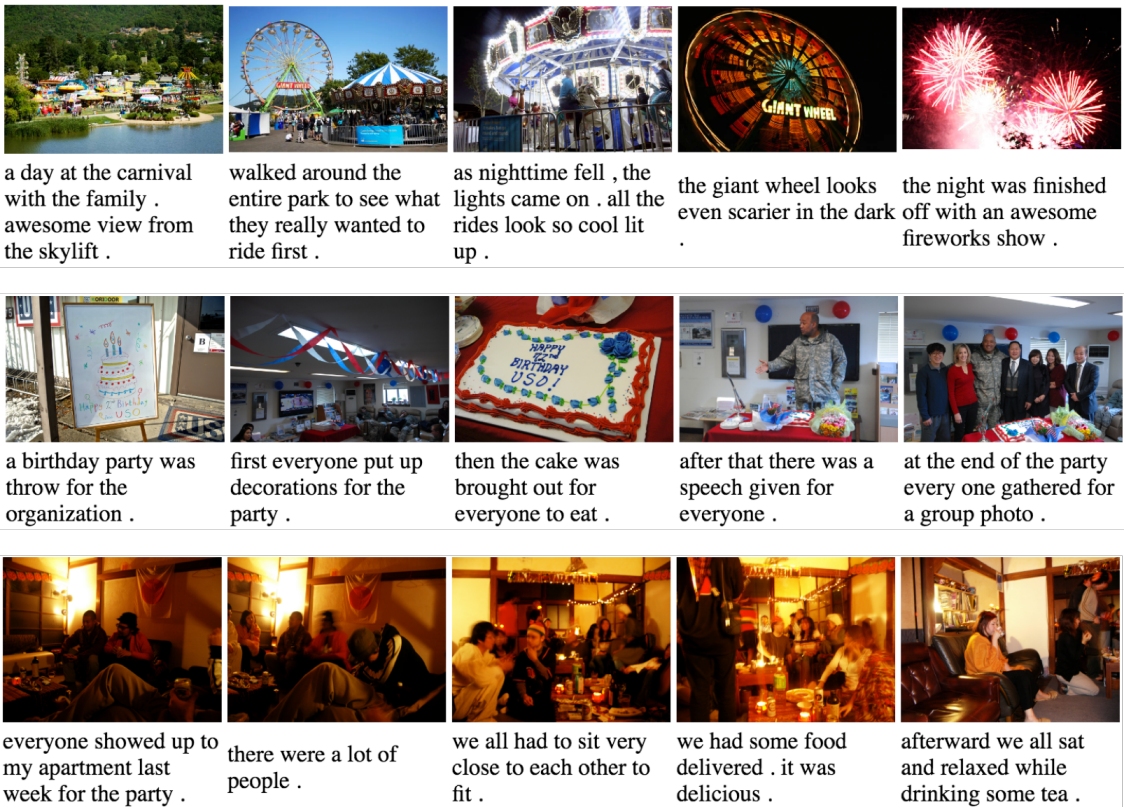
Figure 2: Confusion Matrix for the best performing method i.e Voting ensemble of Pairwise Skip-Thought+CNN and Pairwise Order Neural Position Embedding (NPE).

B Predicted Stories

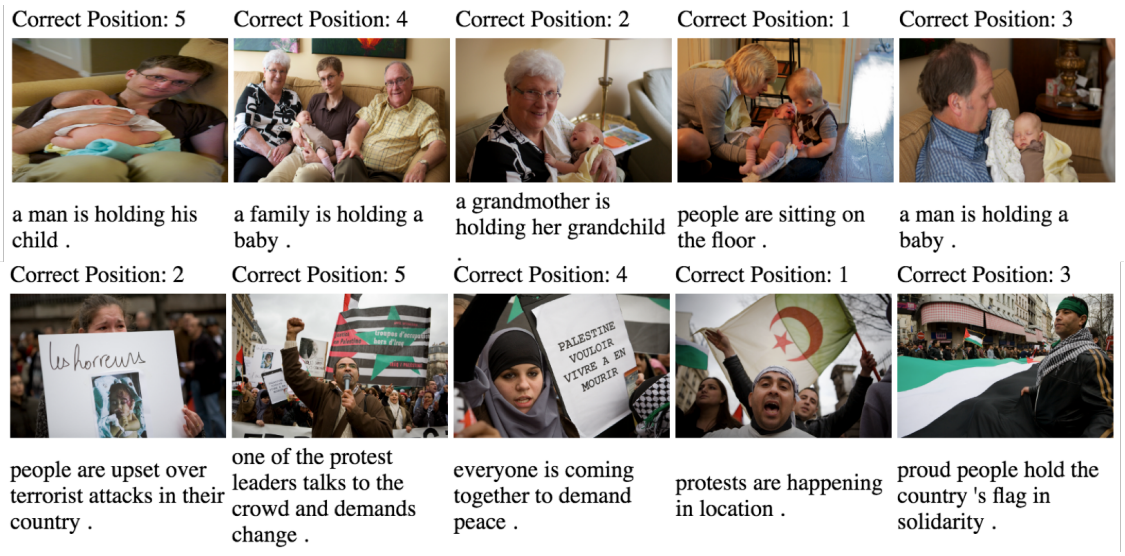
Fig. 3 shows examples of story orders predicted by the best performing model. Fig. 3a shows example stories in which the position of all elements are predicted correctly. We observe that these stories contain language features that may help in determining the order of the elements. Fig. 3b shows stories in which none of the positions are predicted correctly by our model. These two examples show that our model clearly fails when there is no inherent temporal order in the story either via language or images.

C Temporal Common Sense

We visualize our model's *temporal common sense*, in Fig. 4 which consists of word clouds of discriminative words for each position. These words are indicative of the position in which a sentence containing those words would appear.



(a) Qualitative examples of stories whose order was predicted correctly. We can observe that these examples contain language features that give strong signals that may have helped determine the correct order.



(b) Failure examples – stories for which the model failed to predict the correct position of all elements. *Top* - the captions are generic and there seems to be no clear ordering of elements in the story. *Bottom* – again, the story seems to be missing a strong, coherent temporal nature.

Figure 3: Some success and failure examples of story orderings predicted by our best performing model.

(a) First Position (b) Second Position (c) Third Position

(d) Fourth Position (e) Fifth Position

Figure 4: Word cloud corresponding to most discriminative words for each position.

Some of the discriminative words occurring in the first sentence of the story are 'party', 'wedding', etc., probably because our model believes that the start the story describes the setup – the occasion or event. Similarly the indicative words in the second and the third sentences are 'people', 'friend', 'everyone' etc. People often tend to describe meeting friends or family members or going out with them, which seems representative of the middle portion of a short story. As is evident from the discriminative words of the last two positions, our model believes that people tend to conclude the stories using words like 'finally', 'afterwards', tend to talk about 'great day', group 'pictures' with everyone, etc. The example word cloud shows that our model understands to an extent, how a typical story may start, continue, and end, which also usually follows the temporal nature of events.

References

[Basha et al.2012] Tali Basha, Yael Moses, and Shai Avidan. 2012. Photo sequencing. In *Computer Vision–ECCV 2012*, pages 654–667. Springer.

[Boguraev and Ando2005] Branimir Boguraev and Rie Kubota Ando. 2005. Timeml-compliant text analysis for temporal reasoning. In *IJCAI*, volume 5, pages 997–1003.

[Bosselut et al.2016] Antoine Bosselut, Jianfu Chen, David Warren, Hannaneh Hajishirzi, and Yejin Choi. 2016. Learning prototypical event structure from photo albums. In *ACL*.

[Chambers and Jurafsky2008] Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*, volume 94305, pages 789–797. Citeseer.

[Choi et al.2016] Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2016. Video-story composition via plot analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3122–3130.

[Fader et al.2014] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1156–1165. ACM.

[Ferraro et al.2016] Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C Lawrence Zitnick, et al. 2016. Visual storytelling. *arXiv preprint arXiv:1604.03968*.

[Hang2011] LI Hang. 2011. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862.

[Hitzeman et al.1995] Janet Hitzeman, Marc Moens, and Claire Grover. 1995. Algorithms for analysing the temporal structure of discourse. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 253–260. Morgan Kaufmann Publishers Inc.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Kehler2000] Andrew Kehler. 2000. Coherence and the resolution of ellipsis. *Linguistics and Philosophy*, 23(6):533–575.

[Kim and Xing2014] 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.

[Kim et al.2014] Gunhee Kim, Leonid Sigal, and Eric Xing. 2014. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4225–4232.

[Kim et al.2015] Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Joint photo stream and blog post summarization and exploration. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3081–3089. IEEE.

[Kiros et al.2015] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

[Lapata and Lascarides2006] Mirella Lapata and Alex Lascarides. 2006. Learning sentence-internal temporal relations. *J. Artif. Intell. Res.(JAIR)*, 27:85–117.

[Mani and Schiffman2005] Inderjeet Mani and Barry Schiffman. 2005. Temporally anchoring and ordering events in news. *Time and Event Recognition in Natural Language. John Benjamins*.

[Mani et al.2006] Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 753–760. Association for Computational Linguistics.

[Mostafazadeh et al.2016] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. *Proceedings of NAACL HLT, San Diego, California, June. Association for Computational Linguistics*.

[Munkres1957] James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

[Passonneau1988] Rebecca J Passonneau. 1988. A computational model of the semantics of tense and aspect. *Computational Linguistics*, 14(2):44–60.

[Pickup et al.2014] Lyndsey Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William Freeman. 2014. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042.

[Pustejovsky et al.2003] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

[Ren et al.2015] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2935–2943.

[Richardson et al.2013] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 1, page 2.

[Schank and Abelson2013] Roger C Schank and Robert P Abelson. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

[Sigurdsson et al.2016] Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. 2016. Learning visual storylines with skipping recurrent neural networks. *arXiv preprint arXiv:1604.04279*.

[Simonyan and Zisserman2014] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Spearman1904] Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

[Tang et al.2012] 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.

[Trottier1998] David Trottier. 1998. *The screenwriter's bible: A complete guide to writing, formatting, and selling your script*. Silman-James Press.

[Vendrov et al.2015] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

[Wang et al.2016] William Yang Wang, Yashar Mehdad, Dragomir R Radev, and Amanda Stent. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *NAACL*.

[Webber1988] Bonnie Lynn Webber. 1988. Tense as discourse anaphor. *Computational Linguistics*, 14(2):61–73.

[Weston et al.2015] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

[Zhu et al.2015] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and

movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.