Supplementary Material for
Sort Story: Sorting Jumbled Images and Captions into Stories

Harsh Agrawal[*,1] **Arjun Chandrasekaran**[*,1,†]
Dhruv Batra[3,1] **Devi Parikh**[3,1] **Mohit Bansal**[4,2]
[1]Virginia Tech [2]TTI-Chicago [3]Georgia Institute of Technology [4]UNC Chapel Hill
{harsh92, carjun, dbatra, parikh}@vt.edu, mbansal@cs.unc.edu

1 Confusion Matrix for Predicting Position of an Element

We visualize the 5-way classification confusion matrix for our best performing method i.e., Voting ensemble of Pairwise Skip-Thought+Image(CNN) and Pairwise Order (Neural Position Embedding (NPE)) in Fig. 1. The block-diagonal matrix structure shows that the model predicts the first and the last element of a story reasonably well but is often confused by elements in the middle of the story. This visualization suggests that the model has learnt the *three act structure* in stories, i.e., the setup, the middle and the climax.

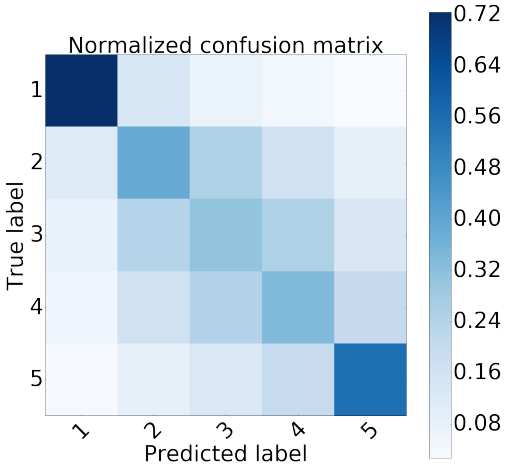


Figure 1: Confusion matrix for predictions from the best performing model i.e Voting ensemble of Pairwise Skip-Thought+image(CNN) and Pairwise Order Neural Position Embedding (NPE).

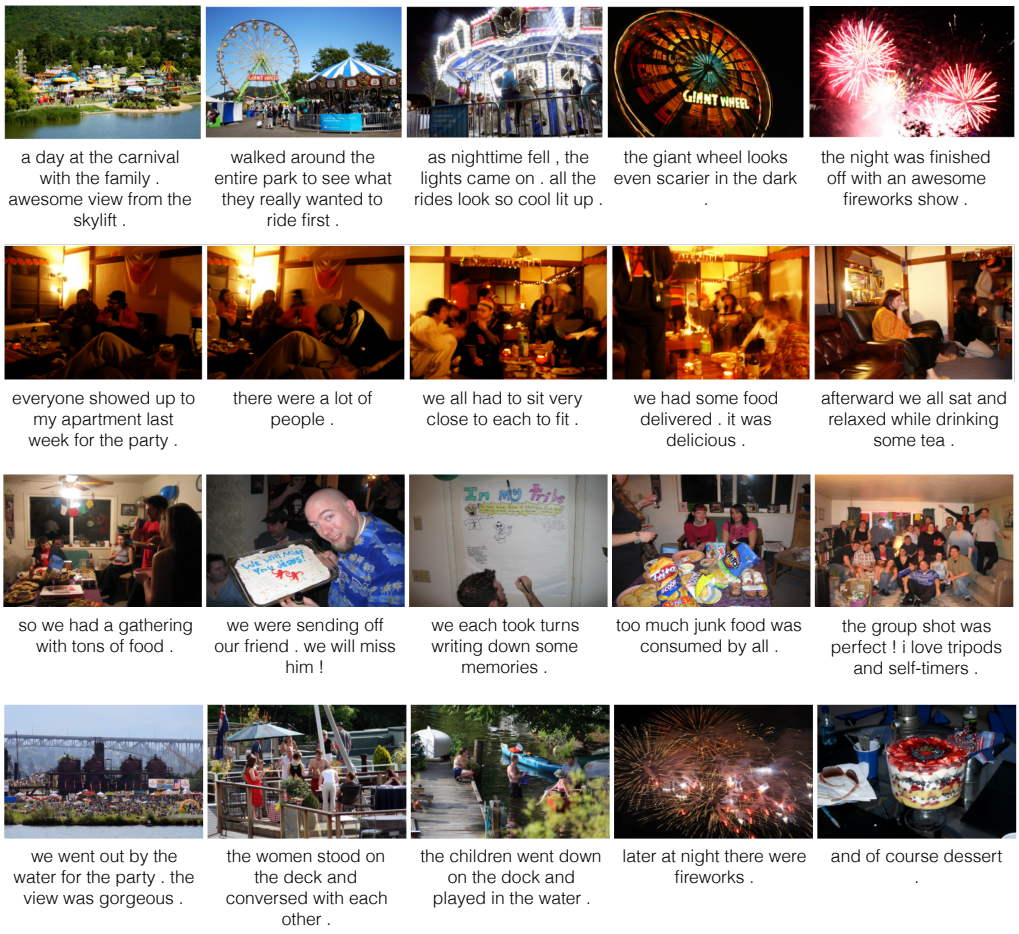
2 Predicted Stories

We present qualitative examples of story orders predicted by the best performing model in Fig. 2. Fig. 2a shows example stories in which the position of all elements are predicted correctly. Fig. 2b 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.

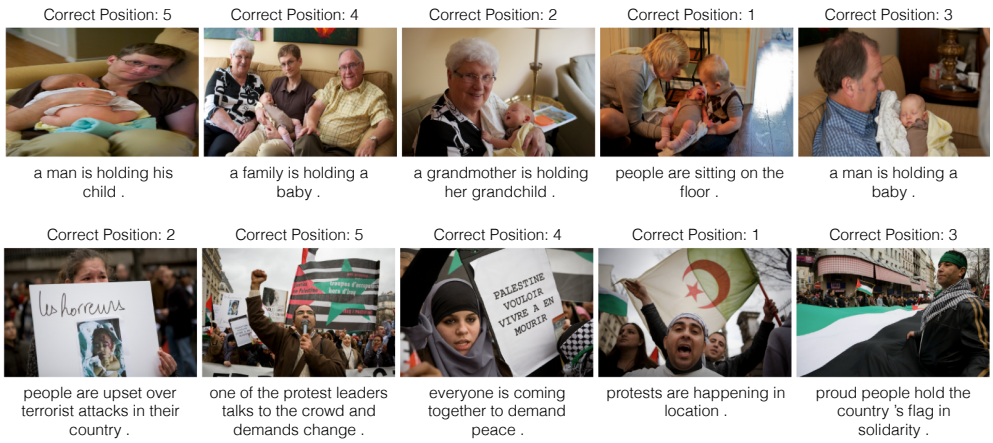
3 Temporal Common Sense

In the word cloud in Fig. 3, we visualize the words that the model finds *discriminative* in correct predictions. These are words from *correctly* predicted stories that the model believes are indicative of sentence positions in a story. The size of a word is proportional to the ratio of its frequency of occurring in that position to other positions. Our model captures events such as 'carnival', 'reunion', and sports topics like 'baseball', 'soccer', 'skate' in the first position. This could be the case because the first sentence of a story usually introduces the event that the story is based on. In Fig. 3e (word-cloud of the last sentence), we also observe that the model correctly learns cue-words such as 'overall', and 'lastly'. It also learns words and events that frequently conclude stories such as 'returned', 'tired', 'winning', 'winner', and 'celebration'.

[*]Denotes equal contribution.
[†]Part of this work was done during an internship at TTIC.



(a) Examples of stories for which the temporal sequence of elements was predicted perfectly.



(b) Examples of stories for which the model failed to predict the correct position of any story element. The elements (images and captions) in a story are generic, with no clear temporal ordering. The stories seem to lack a coherent narrative.

Figure 2: Examples of success and failure cases of temporal order prediction of story elements by our best performing model.

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

(d) Fourth Position (e) Fifth Position

Figure 3: Discriminative words in each position of all correctly predicted stories.