Sort Story: Sorting Jumbled Images and Captions into Stories

Harsh Agrawal*, Arjun Chandrasekaran*, Dhruv Batra*, Devi Parikh†, Mohit Bansal‡

*Virginia Tech, ‡Toyota Technological Institute, †Georgia Institute of Technology, ‡UNC Chapel Hill

**Motivation**

Learn the temporal order of a sequence of events — learn multimodal temporal common sense via images and captions

**Approach**

**Absolute position prediction (Unary)**

- Pre-trained
- LSTM
- MLP
- Share Weight

**Relative position prediction (Pairwise)**

- Pre-trained
- CNN
- MLP
- Share Weight

**Relative position prediction (Neural Position Embeddings — NPE)**

- Pre-trained
- LSTM
- MLP
- Share Weight

**Dataset**

- 10K images and narrative captions form a story
- Train — 40K stories, Val — 5K stories, Test — 5K stories
- We use Microsoft’s SIND (Sequential Image Narrative Dataset)

**Quantitative Results**

- Spearman’s correlation
- Pairwise comparison
- Avg. distance

- Random
- Un SK
- Un SK+Img
- NPE
- PW SK
- PW SK+Img
- PW SK+Img+NPE

**Qualitative Analysis**

- Confusion matrix for predicting position

- ‘First and last story elements are often predicted correctly’
- ‘Loss certainty about the middle elements of the story’

- Model has learnt the 3-act structure

**Temporal common sense**

- Word clouds of discriminative words at each sentence position in the story
- Word size and relative frequency of occurring in that position

**Applications**

- Question answering
- Document summarization
- Human-computer communication

- We use Microsoft’s SIND (Sequential Image Narrative Dataset)
- Image captions are narrative and conversational
- A set of 5 images and associated narrative captions form a story
- Train — 40K stories, Val — 5K stories, Test — 5K stories

**Correct and Incorrect Predictions**

- Correct predictions
- Incorrect predictions

- Correct Positions:
  - Everyone showed up for the party.
  - The view over terrorist attacks in their country.
  - everyone is upset.
  - someone is holding his country.

- Incorrect predictions
  - everyone is upset.
  - someone is holding his country.
  - everyone is upset.
  - someone is holding his country.