

**Overview**

Goal: Learn a bi-directional mapping between images and their sentence-based descriptions

**Usage:**
- Bi-directional retrieval
- Caption generation

**Key Motivation:**
- Visual representations help build long-term memory
- A good caption should capture and help reconstruct the visual representation.

**Evolving visual memory...**

A girl and boy knocked down a tower.

**Background**

Previous RNN Model:

\[
P(w_s|V,W_{t-1}) = P(w_s|V_{t-1})P(V|W_{t-1})
\]

Training:
- Per stage model, every step tries to reconstruct the image
- Weight update from visual memory to image is performed from end to start

Retrieval:
- Given a sentence, evaluate the likelihood that it can be generated by using each image as an input
- Image to sentence retrieval is normalized by sentence length
- Using visual memory helps the performance

Generation:
- First sample sentence length from a prior
- With fixed length, sample the most likely caption

**Conclusions**

- Explicit visual memory is helpful
- Visual memory can be learned even with a single image per sentence
- Simple RNNs can remember long-term concepts
- Model is decomposable for bi-directional generation

**Results**

- Human Evaluation
  - 5.3% of our captions (Our Approach + VGG) are preferred to human captions, and 15.9% of equal quality

- **Caption generation**
  - A good caption should capture and help reconstruct the visual representation.

- **Visual Feature Reconstruction**
  - A long-term visual memory

- **Caption Generation**
  - First sample sentence length from a prior
  - With fixed length, sample the most likely caption

- **Our Model**
  - \[ P(w_s, V|W_{t-1}) = P(w_s|V|W_{t-1})P(V|W_{t-1}) \]
  - Bi-directional retrieval
  - Caption generation

- **Visual Feature Reconstruction**
  - A good caption should capture and help reconstruct the visual representation.

- **Background**
  - Previous RNN Model:
    \[
P(w_s|V,W_{t-1}) = P(w_s|V_{t-1})P(V|W_{t-1})
\]
  - Training:
    - Per stage model, every step tries to reconstruct the image
    - Weight update from visual memory to image is performed from end to start
  - Retrieval:
    - Given a sentence, evaluate the likelihood that it can be generated by using each image as an input
    - Image to sentence retrieval is normalized by sentence length
    - Using visual memory helps the performance
  - Generation:
    - First sample sentence length from a prior
    - With fixed length, sample the most likely caption

- **Conclusions**
  - Explicit visual memory is helpful
  - Visual memory can be learned even with a single image per sentence
  - Simple RNNs can remember long-term concepts
  - Model is decomposable for bi-directional generation