Towards Video Captioning with Spatio-Temporal Hierarchical Attention Networks

Spencer Whitehead
11/9/2017
Outline

● Problem Definition
● Recurrent Neural Networks
● Convolutional Neural Networks
● Attention Mechanisms
  ○ Encoder-Decoder Model (Cho et al., 2014)
  ○ Attention Mechanism (Bahdanau et al., 2015)
  ○ Image Captioning (Xu et al., 2015)
● Previous Approaches
● Spatio-Temporal Hierarchical Attention Network
● Initial Results
Problem Definition
Problem Definition

- Given a video, generate a natural language description of objects activities depicted in the video.
  - Typically a single sentence for a specific clip
Recurrent Neural Networks
Recurrent Neural Networks

- Long Short-Term Memory Network (LSTM)
- Input can be word embedding
- Output predicts probability distribution over vocabulary
Convolutional Neural Networks
Convolutional Neural Networks

Krizhevsky et al., 2012

*fc7 layer
Attention Mechanisms
Attention Mechanisms

- Based upon Encoder-Decoder model
- Breakthrough in machine translation (MT)
- Allows system to selectively focus on certain portions of source sentence
- Improved state-of-the-art
Encoder-Decoder Model

- Introduced by Cho et al., 2014
- Two RNNs: 1 Encoder, 1 Decoder
- Encoder scans over input to get fixed vector representation
- Decoder uses Encoder representation as initial state

Figure 1: An illustration of the proposed RNN Encoder–Decoder.

Cho et al., 2014
Attention Mechanism

- Introduced by Bahdanau et al., 2015 and based upon Encoder-Decoder Framework
- Multi-layer perceptron
  - Takes in hidden state of decoder at previous step, output of previous step, and hidden state of encoder
  - At each step of decoding, attention mechanism computes a score for the encoder hidden states
  - Take a weighted average of scored encoder hidden states to produce a context vector.
  - Decoder generates next output conditioned on context vector, previous decoder hidden state, and previous output

Bahdanau et al., 2015
Image Captioning

- Can use CNN encoder
- Without attention (Vinyals et al., 2015) and with visual attention (Xu et al., 2015)
Previous Approaches

- **Early work:**
  - Two-stage pipeline:
    - i. Extract a subject, verb, object (SVO) triple from video
    - ii. Template generation to create sentence
  - Guadarrama et al., 2013; Thomason et al., 2014; Huang et al. 2013

- **Neural approaches**
  - Mean pool static frame level visual features (CNN) of the video and input mean pool features into language decoder (LSTM) (Venugopalan et al., 2015b)
  - Sequence-to-sequence model with video encoder and language decoder LSTMs (Venugopalan et al., 2015a)
  - Incorporate linguistic knowledge by fusing caption decoder with language model which uses pre-trained word embeddings (Venugopalan et al., 2016)
Previous Approaches

- **With Attention:**
  - Add attention mechanisms between encoder neural network and decoder neural network to learn temporal relations between video frames and caption words (Yao et al., 2015, Pan et al., 2016)

- **Hierarchical RNN:**
  - Two-level hierarchical RNN encoder (Pan et al., 2016)
    - RNNs iterate over output of RNNs lower in hierarchy
  - Hierarchical RNN that generates sentences at the first level and then paragraphs at the second level (Yu et al., 2016)
    - Uses joint spatial/temporal attention by applying CNN to subsections of image and taking fc7 response as representation for subsection
    - In captioning, they simply use fc7 layer response as frame representation
Video Captioning

- Generate a sentence description given a video
- Can utilize encoder decoder framework without attention

Venugopalan et al., 2015
Previous Approaches

- These approaches tend to ignore spatial information contained in the frames by only using the fc6/fc7 layer as a representation for each frame
  - This has been shown to be beneficial by Xu et al. (2015) for image captioning
- Motion has been captured in previous approaches
  - However, Yao et al. (2015) simply concatenate 3D CNN features (which capture local temporal structure) and attend to only the global temporal structure
    - This makes no connection between the visual features of each frame and the local temporal structure
- Multi-Task Video Captioning with Video and Entailment Generation (Pasunuru and Bansal, 2017) *Outstanding paper award at ACL 2017
  - Performs multitask learning (video captioning, unsupervised video prediction, and entailment generation) with different shared components for each task to enhance video captioning
  - Baseline that is used, feeds feature vector from each frame into BiLSTM encoder and LSTM decoder with attention
Temporal Attention

- Introduced by Yao et al., 2015
- Attention on local and global temporal structure
Multi-Task Video Captioning

- Pasunuru and Bansal, 2017 use Encoder-Decoder with temporal attention as baseline
- Multi-task learning for improvements

Multi-task learning from Pasunuru and Bansal, 2017
Spatio-Temporal Hierarchical Attention Network
Spatio-Temporal Hierarchical Attention Network

1. Input video, \( V = \{v_1, ..., v_n\} \), \( v_i \) is a frame
2. Frame feature extraction (CNN)
   - For each frame, use a CNN to extract visual features vectors from each section of the frame (e.g. top-left corner, top-right corner, ...)
   - Input: \( V \)
   - Output: for each \( v_i \), \( a_i = \{a_{i1}, ..., a_{iL}\} \), \( a_{ij} \) in \( \mathbb{R}^D \)
     - \( A = \{a_1, ..., a_n\} \)
   - \( a_{ij} \) is a representation of the \( j^{th} \) location (covering some section of image, not individual pixel) in frame \( v_i \)
Spatio-Temporal Hierarchical Attention Network

3. Spatial Attention
   ○ For each frame, compute a weighted average of the location vectors so certain sections of the frame are emphasized more than others
   ○ This yields a single, spatial context vector representing the frame
   ○ Input: for each \( v_i \), \( a_i=\{a_{i1},...,a_{iL}\} \)
   ○ For each \( a_{ij} \):
     ■ \( e_{ij} = f_{s\text{-att}}(a_{ij}, h_{i-1}) \), where \( f_{s\text{-att}} \) is single layer MLP and \( h_i \) is temporal encoder hidden state (if temporal encoder is BiLSTM, \( e_{ij} = f_{s\text{-att}}(a_{ij}) \))
     ■ \( \xi_{ij} = \text{softmax}(e_{ij}) \)
   ○ Output: For each frame \( v_i \), \( z_i = \sum_j^L \xi_{ij} a_{ij} \)
Spatio-Temporal Hierarchical Attention Network

4. Temporal Encoder (either LSTM or Bidirectional LSTM)
   - Given the spatial representation of each frame, learn a temporal representation using an LSTM or BiLSTM
   - The resulting representation contains information about the most important parts of the images (objects/scenery) as well as the temporal relations of these important parts between frames
   - Input: $z_i$, $i=1,...,n$
   - For each $z_i$:
     - LSTM iterates over the sentence
       - $h_i = \text{LSTM}(z_i)$, $i = 1,...,n$
     - BiLSTM iterates over the sentence forward (fw) and backward (bw)
       - $h_i^{(fw)} = \text{LSTM}_{fw}(z_i)$, $i = 1,...,n$
       - $h_i^{(bw)} = \text{LSTM}_{bw}(z_i)$, $i = n,...,1$
       - $h_i = [h_i^{(fw)}, h_i^{(bw)}]$
   - Output: $h_i$, $i = 1,...,n$
Spatio-Temporal Hierarchical Attention Network

5. Temporal Attention
   - Given the spatio-temporal representations of the frames, at each step of generation (see temporal decoder), compute a weighted average of these representations so the most important frames (sections of the video) are emphasized.
   - This yields a single, spatio-temporal context vector representing the most important parts of the most important frames.
   - Input: $h_i$, $i = 1, ..., n$
   - For each $h_i$:
     - $e_i = f_{t-att}(h_i, s_{t-1})$, where $f_{t-att}$ is single layer MLP and $s_t$ is decoder hidden state.
     - $\alpha_i = \text{softmax}(e_i)$
   - Output: $c_t = \sum_j \alpha_j h_j$
Spatio-Temporal Hierarchical Attention Network

6. Language decoder
   ○ Use LSTM language model to compute hidden state to generate words
   ○ At each step of generation, the hidden state is conditioned on the previous
     word, previous hidden state, and the spatio-temporal context vector computed
     by the temporal attention
   ○ Input: $c_t, w_{t-1}, s_{t-1}$
   ○ At each time $t$, compute
     $$s_t = \text{LSTM}(s_{t-1}, w_{t-1}, c_t)$$
   ○ Output: At each time $t$, $s_t$

7. Deep output layer (Pascanu et al., 2014)
   ○ Input: $Ew_{t-1}, s_t, c_t$
     $$\text{Compute } p(w_t|A, w_1^{t-1}) \propto \exp(L_o(Ew_{t-1} + L_h s_t + L_c c_t))$$
   ○ Output: $p(w_t|A, w_1^{t-1})$

8. Output word sampling
   ○ Sample from $p(w_t|A, w_1^{t-1})$
   ○ Can take argmax($w_i$), which yields index of max probability word in vocabulary
Initial Results
Qualitative Examples (Good)

- A woman is putting on makeup.
- A man is riding a bike.
Qualitative Examples (Good) Attention Visualization

- Temporal attention appears to group relevant words together and more meaningful words have more evenly distributed attention
  - Ex: A woman is putting on makeup.
  - Temporal attention emphasis:
    - ‘A’: frame 22
    - ‘woman’: frame 67
    - ‘is’: frame 67
    - ‘putting’: frame 48
    - ‘on’: frame 48
    - ‘makeup’: frame 79
Qualitative Examples (Good) Attention Visualization

- Spatial attention does not always align with meaningful content
  - Ex: A man is riding a bike.
  - Temporal attention emphasis:
    - ‘A’: frame 61
    - ‘man’: frame 71
    - ‘is’: frame 56
    - ‘riding’: frame 71
    - ‘a’: frame 56
    - ‘bike’: frame 56
Qualitative Examples (Bad)

- A boy is talking and eating.
- A man is cutting a slices of carrot.
Qualitative Examples (Bad) Attention Visualization

- Sentence realization is decent, but still has flaws.
- Training sentences are rather homogeneous
  - Most follow form: ‘A <generic entity> is <present tense verb>...’
- Ex: A man is cutting a slices of carrot.
Qualitative Examples (Bad) Attention Visualization

- Worse generated sentences tend to focus on initial frame
  - Ex: A boy is talking and eating.
  - Temporal attention emphasis:
    - ‘A’: frame 48
    - ‘boy’: frame 45
    - ‘is’: frame 0
    - ‘talking’: frame 0
    - ‘and’: frame 0
    - ‘eating’: frame 52
Qualitative Examples (Ugly)

- A man is walking a horse.
Qualitative Examples (Ugly) Attention Visualization

- Worse generated sentences tend to focus on initial frame
  - Ex: A man is walking a horse.
  - Temporal attention emphasis:
    - ‘A’: frame 7
    - ‘man’: frame 8
    - ‘is’: frame 0
    - ‘walking’: frame 0
    - ‘a’: frame 0
    - ‘horse’: frame 0
Outline

● Problem Definition
● Recurrent Neural Networks
● Convolutional Neural Networks
● Attention Mechanisms
  ○ Encoder-Decoder Model (Cho et al., 2014)
  ○ Attention Mechanism (Bahdanau et al., 2015)
  ○ Image Captioning (Xu et al., 2015)
● Previous Approaches
● Spatio-Temporal Hierarchical Attention Network
● Initial Results
Questions?
Thank You!
References

- Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In CVPR.
- Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond J Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In COLING.
References

- Subhashini Venugopalan, Lisa Anne Hendricks, Raymond Mooney, and Kate Saenko. 2016. Improving lstm-based video description with linguistic knowledge mined from text. In EMNLP.
- Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015a. Sequence to sequence-video to text. In ICCV.
- Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In CVPR.
References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR.
- Oriol Vinyals et al. 2015. Show and tell: A neural image caption generator. In CVPR.
- Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In CVPR.
- Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In CVPR.