Vision Guided Language Generation

Min Sun

National Tsing Hua University
Overview

• Captioning
  – For Image
  – Referring Expression
  – For Video

• Question Answering
  – For Image
  – For Video

• Others
  – Storytelling
  – Visual-aware HCI
Captioning

• I have a **CNN**, I have a **RNN** -> **Novel Sentences**

Captioning

Sequences in the Output

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015
Captioning

Two-layer word embedding system

The m-RNN model for one time frame

(b). The m-RNN model

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). ICLR 2015
Captioning - **Attention**

- **Attention** mechanism: per word attention.

[Diagram showing the process of image caption generation with attention]

Captioning - Attention

- Attention mechanism: per word attention.

Soft-attention, deterministic, Backpropagation

Hard-attention, stochastic, lower-bound or RL

Captioning - Attributes

- **Semantic Attention** (e.g., surfboard, etc.)

Captioning - *Attributes*

- **Semantic** Attention (e.g., surfboard, etc.)

Captioning - Attributes

• Great performance on COCO captioning

Attributes: [bananas: 1] [market: 0.995] [bunch: 0.553] [table: 0.51] [flowers: 0.454] [people: 0.431] [yellow: 0.377]

LSTM: a group of people standing around a market.
A-LSTM: a group of people standing around a bunch of bananas.

BOOSTING IMAGE CAPTIONING WITH ATTRIBUTES
Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, Tao Mei. ICLR 2017 under review
Captioning – *Rich Caption*

- Landmark and Celebrity

“Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.”

Kenneth Tran et al., Rich Image Captioning in the Wild. CVPR 2016
Captioning – *Curriculum Learning*

• gently change the training process from a fully guided scheme using the true previous token, towards a less guided scheme which mostly uses the generated token instead

Captioning – *Unpaired Data*

- generating descriptions of novel objects which are not present in paired image-sentence datasets

Captioning – *Unpaired Data*

Captioning – *DenseCap*

- Both localize and describe salient regions in images in natural language.

Justin Johnson, Andrej Karpathy, Li Fei-Fei, *DenseCap: Fully Convolutional Localization Networks for Dense Captioning*, CVPR 2016
Captioning – *DenseCap*

- Caption and localization layers end-to-end trainable

Justin Johnson, Andrej Karpathy, Li Fei-Fei, DenseCap: Fully Convolutional Localization Networks for Dense Captioning, CVPR 2016
Captioning – *Object Descriptions*

- Generation and Comprehension of Unambiguous OD

Captioning – **Object Retrieval**

- using Natural Language

\[ \text{score}_\text{box} = p(S=\text{`man in middle with blue shirt and blue shorts`} \mid \text{\_box, \_im, \_spatial}) \]

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, Trevor Darrell. Natural Language Object Retrieval CVPR 2016
Captioning – *Referring Expression*

- Joint inference among all objects

Captioning – *Referring Expression*

• Context Between Objects for Referring Expression

Varun K. Nagaraja Vlad I. Morariu Larry S. Davis, Modeling Context Between Objects for Referring Expression Understanding, ECCV 2016
Captioning – Dataset

• Peter Young, Alice Lai, Micah Hodosh, Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. (*Flickr30K*) TCAL 2014

• Tsung-Yi Lin et al., *Microsoft COCO*: Common Objects in Context. ECCV 2014

Challenge

Image captioning

- **Leaderboard** of MS COCO image captioning
  - Rank 1 in both external and internal ranking lists, in terms of all performance metrics (July 21)
- COCO dataset
  - 123,287 images (82,783 for training + 40,504 for validation)
  - 5 sentences per image (AMT workers)
I have a RNN-Encode, I have a RNN-Decoder: Video-to-Text

Video Captioning

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko. Sequence to Sequence – Video to Text. ICCV'15
Video Captioning-Attention

- Frame-level soft-attention

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, Aaron Courville, Describing Videos by Exploiting Temporal Structure. ICCV 2015
Video Captioning-Joint Embedding

- Additional Relevance Loss

Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, Yong Rui. Jointly Modeling Embedding and Translation to Bridge Video and Language, CVPR 2016
Exploiting video temporal structure in a longer range

Pingbo Pan et al., Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning. CVPR 2016
Video Captioning-Generate a Paragraph

- A sentence generator and a paragraph generator
- Spatial- and Temporal-Attention
- Paragraph state to initialize Recurrent I

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, Wei Xu, Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks. CVPR 2016
Video Captioning – Title (Highlight)

Title (most salient event): Bmx rider gets *hit by scooter* at park

Captions: A man riding on bike. A man does a stunt on a bmx bike.

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, Min Sun. Title Generation for User Generated Videos. ECCV 2016
Video Captioning—Bi-direction

Yi Bin et al., Bidirectional Long-Short Term Memory for Video Description. ACM MM 2016
Video Captioning – Dataset

• Rohrbach et al. MPII Movie Description (MPII-MD). CVPR 2016
• Torabi et al. Montreal Video Annotation Dataset (M-VAD). Arxiv 2016
• Jun Xu, Tao Mei, Ting Yao and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. CVPR 2016
• Zhen et al. Video Title in the Wild (VTW). ECCV 2016
Microsoft Video to Language Challenge

77 teams registered challenge
22 teams submitted results
Awards will be announced at ACM MMM
Question Answering
Question Answering

• RNN to encode a question and Image; RNN to decode an answer (multiple words); Single-RNN

Mateusz Malinowski, Marcus Rohrbach, Mario Fritz, Ask Your Neurons: A Neural-based Approach to Answering Questions about Images, ICCV 2015
Question Answering

- limited answer space for easy evaluation

Mengye Ren, Ryan Kiros, Richard Zemel, Exploring Models and Data for Image Question Answering, ICML 2015
Question Answering

• Separate LSTM-Q and LSTM-A

Question Answering

• Point-wise multiplication

“How many horses are in this image?”

Question Answering – Attention

- Stack Attention Network (SAN)

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, Stacked Attention Networks for Image Question Answering, CVPR 2016
Question Answering – Attention

- Dynamic Memory Network

Question Answering – Attention

- word to patch at 1\textsuperscript{st} hop; whole Q at 2\textsuperscript{nd} hop.

Huijuan Xu, Kate Saenko, Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering, ECCV 2016
Question Answering – Attention

• Averaged representation of word2vec vectors for language

Kevin J. Shih, Saurabh Singh, Derek Hoiem, Where To Look: Focus Regions for Visual Question Answering, CVPR 2016
Question Answering – Attention

• Dual Attention

Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. CVPR'16 VQA Challenge Workshop
Question Answering – Attention

- Dual Attention

Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. CVPR'16 VQA Challenge Workshop
Question Answering – Dynamic Parameter

- Dynamic Parameter Layer (hashing)

Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han, Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction, CVPR 2016
Question Answering – Bilinear Pooling

• Outer product of the visual and textual vectors

Question Answering – Knowledge

• External Knowledge

Internal Textual Representation:
A group of people enjoying a sunny day at the beach with umbrellas in the sand.

External Knowledge:
An umbrella is a canopy designed to protect against rain or sunlight. Larger umbrellas are often used as points of shade on a sunny beach. A beach is a landform along the coast of an ocean. It usually consists of loose particles, such as sand....

Question Answering:
Q: Why do they have umbrellas? A: Shade.

Question Answering – Knowledge

• External Knowledge

Question Answering – Dataset

• DAQUAR – Malinowski and Fritz. NIPS 2014
• VQA - based on MSCOCO images. ICCV 2015
• COCO-QA - based on MSCOCO images. Ren et al. ICML 2015
• Yuke Zhu, Oliver Groth, Michael Bernstein, Li Fei-Fei, Visual7W: Grounded Question Answering in Images, CVPR 2016.
Challenge

VQA Real Image Challenge (Multiple-Choice)
Organized by vqateam - Current server time: Nov. 24, 2016, 5:16 a.m. UTC

Current
Real test2015 (mc)
Oct. 21, 2015, midnight UTC

Next
Real test2015 (mc)
Oct. 21, 2015, midnight UTC
Video Question Answering

- Learning to rank multiple choices

Video Question Answering

• Multiple Extensions for Video-QA

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, Min Sun. Leveraging Video Descriptions to Learn Video Question Answering. AAAI 2017
Video Question Answering – Dataset

• VTW Video-QA dataset

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, Min Sun. Leveraging Video Descriptions to Learn Video Question Answering. AAAI 2017
Video Question Generation

- How many horses are in the field? ✗
- Who won the race? ✓

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, Lucy Vanderwende, Generating Natural Questions About an Image, ACL 2016
Video Question Answering – Dataset

- Moive-QA

Others
Storytelling – Retrieval

- **Retrieve** fluent sequential multiple sentences

---

Cesc Chunseong Park, Gunhee Kim. Expressing an Image Stream with a Sequence of Natural Sentences. NIPS 2015
Storytelling – Generate

- Sequential Images Narrative Dataset (SIND)

| DII | SIS |
| --- | --- | --- |
| A group of people that are sitting next to each other. | Having a good time bonding and talking. | Sky illuminated with a brilliance of gold and orange hues. |
| Adult male wearing sunglasses lying down on black pavement. | [M] got exhausted by the heat. | The sun is setting over the ocean and mountains. |

**Figure 1:** Example language difference between descriptions for images in isolation (DII) vs. stories for images in sequence (SIS).

Ting-Hao (Kenneth) Huang et al., Visual Storytelling, NAACL 2016
Video-Commenting

- Deep Multi-View Embedding Model

**Input Video:**

**Output Comment:**
- Motivated me to go beyond my limits in skateboarding!!!

**Human Made Comment:**
- He should be a new character in the next skateboarding game.

**Output Sentence:**
- A man is doing a trick on a skateboarding.

Yehao Li, Ting Yao, Tao Mei, Hongyang Chao, Yong Rui, “Share-and-Chat: Achieving Human-Level Video Commenting by Search and Multi-View Embedding,” ACM Multimedia (MM), 2016
Video-Chatbot

• human-level emotional comments

Future

• Vision Guided Language-based HCI
  – Chatbot
  – Smart Assistant

• Storytelling
  – Public events (e.g., newspaper)
  – Personal events (e.g., personal blog post)
Thanks!