handong1587 / **handong1587.github.io**

Branch: **master** ▾ **handong1587.github.io** / _posts / deep_learning / **2015-10-09-captioning.md** Find file Copy path

handong1587 add arxiv paper d26b234 10 days ago

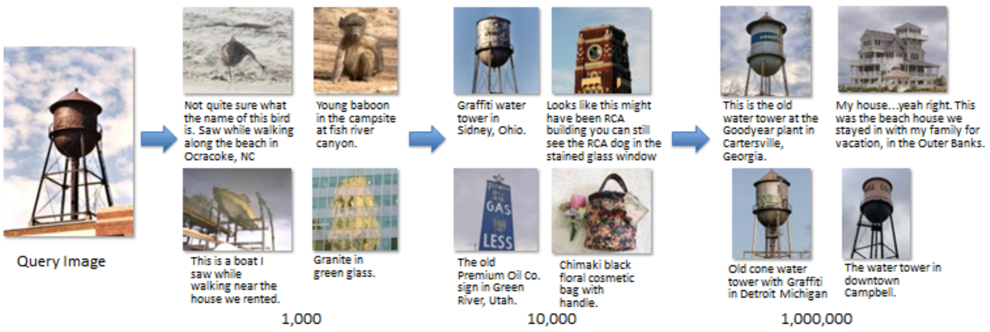
1 contributor

472 lines (288 sloc) 22.6 KB

layout	category	title	date
post	deep_learning	Image / Video Captioning	2015-10-09

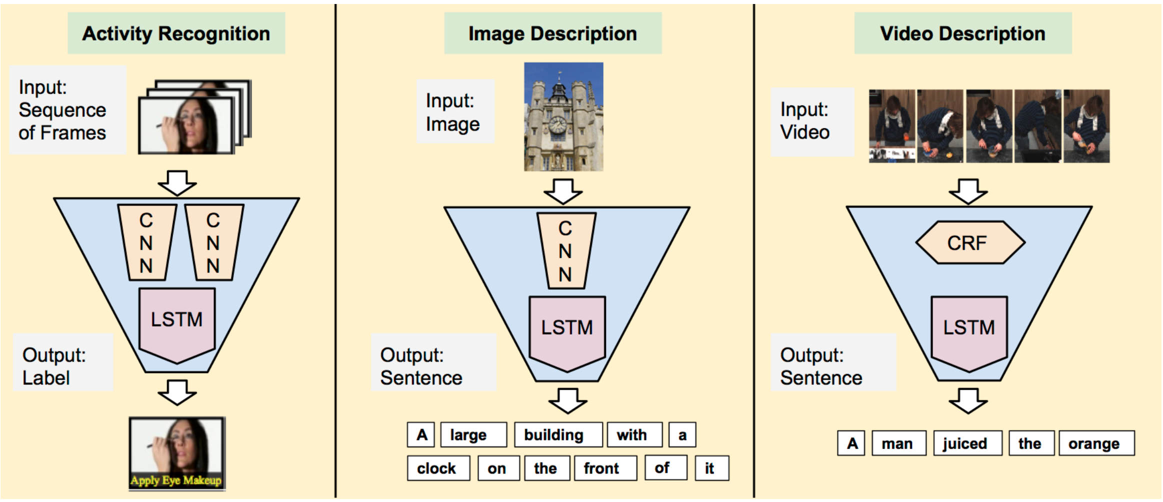
Papers

Im2Text: Describing Images Using 1 Million Captioned Photographs


Query Image
Not quite sure what the name of this bird is. Saw while walking along the beach in Ocracoke, NC
Young baboon in the campsite at fish river canyon.
This is a boat I saw while walking near the house we rented.
Granite in green glass.
1,000
Graffiti water tower in Sidney, Ohio.
Looks like this might have been RCA building you can still see the RCA dog in the stained glass window
The old Premium Oil Co. sign in Green River, Utah.
Chimaki black floral cosmetic bag with handle.
10,000
This is the old water tower at the Goodyear plant in Cartersville, Georgia.
My house...yeah right. This was the beach house we stayed in with my family for vacation, in the Outer Banks.
Old cone water tower with Graffiti in Detroit Michigan
The water tower in downtown Campbell.
1,000,000

- paper: http://tamaraberg.com/papers/generation_nips2011.pdf
- project: http://vision.cs.stonybrook.edu/~vicente/sbucaptions/

Long-term Recurrent Convolutional Networks for Visual Recognition and Description



- intro: Oral presentation at CVPR 2015. LRCN
- project page: http://jeffdonahue.com/lrcn/

- arxiv: http://arxiv.org/abs/1411.4389
- github: https://github.com/BVLC/caffe/pull/2033

Show and Tell

Show and Tell: A Neural Image Caption Generator

- intro: Google
- arxiv: http://arxiv.org/abs/1411.4555
- github: https://github.com/karpathy/neuraltalk
- gitxiv: http://gitxiv.com/posts/7nofxjoYBXga5XjtL/show-and-tell-a-neural-image-caption-nic-generator
- github: https://github.com/apple2373/chainer_caption_generation
- github(TensorFlow): https://github.com/tensorflow/models/tree/master/im2txt
- github(TensorFlow): https://github.com/zsdonghao/Image-Captioning

Image caption generation by CNN and LSTM

↑ a living room with a couch and a television
↑ a man riding a bike on a beach
a man is walking down the street with a suitcase ↗

- blog: http://t-satoshi.blogspot.com/2015/12/image-caption-generation-by-cnn-and-lstm.html
- github: https://github.com/jazzsaxmafia/show_and_tell.tensorflow

Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge

- arxiv: http://arxiv.org/abs/1609.06647
- github: https://github.com/tensorflow/models/tree/master/im2txt

Learning a Recurrent Visual Representation for Image Caption Generation

- arxiv: http://arxiv.org/abs/1411.5654

Mind's Eye: A Recurrent Visual Representation for Image Caption Generation

- intro: CVPR 2015
- paper: http://www.cs.cmu.edu/~xinleic/papers/cvpr15_rnn.pdf

Deep Visual-Semantic Alignments for Generating Image Descriptions

- intro: "propose a multimodal deep network that aligns various interesting regions of the image, represented using a CNN feature, with associated words. The learned correspondences are then used to train a bi-directional RNN. This model is able, not only to generate descriptions for images, but also to localize different segments of the sentence to their corresponding image regions."
- project page: http://cs.stanford.edu/people/karpathy/deepimagesent/
- arxiv: http://arxiv.org/abs/1412.2306
- slides: http://www.cs.toronto.edu/~vendrov/DeepVisualSemanticAlignments_Class_Presentation.pdf
- github: https://github.com/karpathy/neuraltalk
- demo: http://cs.stanford.edu/people/karpathy/deepimagesent/rankingdemo/

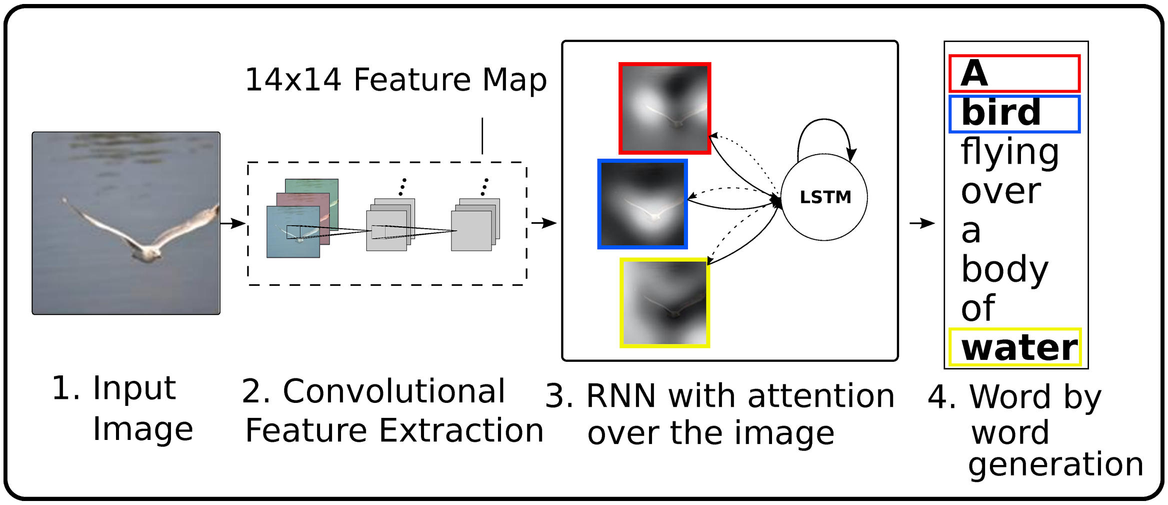
Deep Captioning with Multimodal Recurrent Neural Networks

- intro: m-RNN. ICLR 2015

- intro: "combines the functionalities of the CNN and RNN by introducing a new multimodal layer, after the embedding and recurrent layers of the RNN."
- homepage: http://www.stat.ucla.edu/~junhua.mao/m-RNN.html
- arxiv: http://arxiv.org/abs/1412.6632
- github: https://github.com/mjhucla/mRNN-CR
- github: https://github.com/mjhucla/TF-mRNN

Show, Attend and Tell

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (ICML 2015)


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

- project page: http://kelvinxu.github.io/projects/capgen.html
- arxiv: http://arxiv.org/abs/1502.03044
- github: https://github.com/kelvinxu/arctic-captions
- github: https://github.com/jazzsaxmafia/show_attend_and_tell.tensorflow
- github(TensorFlow): https://github.com/yunjey/show-attend-and-tell-tensorflow
- demo: http://www.cs.toronto.edu/~rkiros/abstract_captions.html

Automatically describing historic photographs

- website: https://staff.fnwi.uva.nl/d.elliott/loc/

Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images

- arxiv: http://arxiv.org/abs/1504.06692
- homepage: http://www.stat.ucla.edu/~junhua.mao/projects/child_learning.html
- github: https://github.com/mjhucla/NVC-Dataset

What value do explicit high level concepts have in vision to language problems?

- arxiv: http://arxiv.org/abs/1506.01144

Aligning where to see and what to tell: image caption with region-based attention and scene factorization

- arxiv: http://arxiv.org/abs/1506.06272

Learning FRAME Models Using CNN Filters for Knowledge Visualization (CVPR 2015)

- project page: http://www.stat.ucla.edu/~yang.lu/project/deepFrame/main.html

- arxiv: http://arxiv.org/abs/1509.08379
- code+data: http://www.stat.ucla.edu/~yang.lu/project/deepFrame/doc/deepFRAME_1.1.zip

Generating Images from Captions with Attention

- arxiv: http://arxiv.org/abs/1511.02793
- github: https://github.com/emansim/text2image
- demo: http://www.cs.toronto.edu/~emansim/cap2im.html

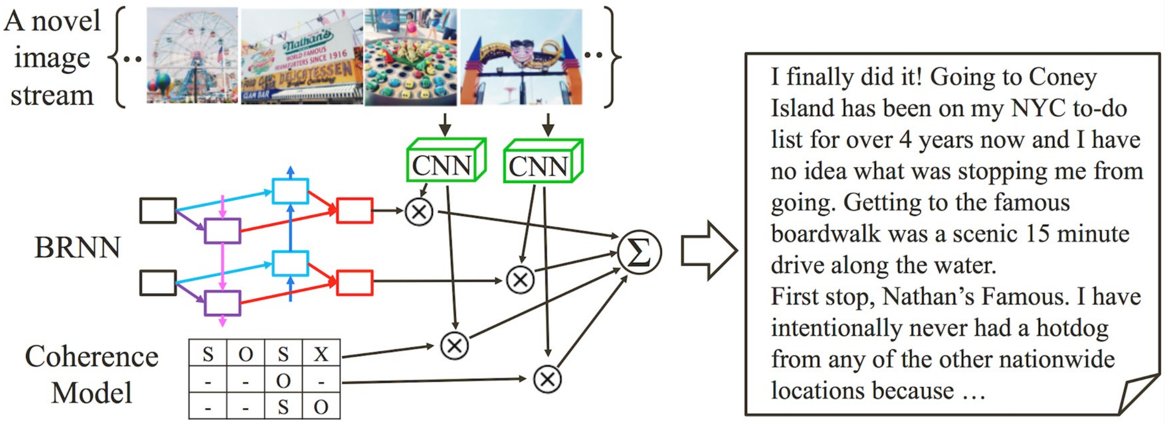
Order-Embeddings of Images and Language

- arxiv: http://arxiv.org/abs/1511.06361
- github: https://github.com/ivendrov/order-embedding

DenseCap: Fully Convolutional Localization Networks for Dense Captioning

- project page: http://cs.stanford.edu/people/karpathy/densecap/
- arxiv: http://arxiv.org/abs/1511.07571
- github(Torch): https://github.com/jcjohnson/densecap

Expressing an Image Stream with a Sequence of Natural Sentences



S	O	S	X
-	-	O	-
-	-	S	O

I finally did it! Going to Coney Island has been on my NYC to-do list for over 4 years now and I have no idea what was stopping me from going. Getting to the famous boardwalk was a scenic 15 minute drive along the water.
First stop, Nathan's Famous. I have intentionally never had a hotdog from any of the other nationwide locations because …

- intro: NIPS 2015. CRCN
- nips-page: http://papers.nips.cc/paper/5776-expressing-an-image-stream-with-a-sequence-of-natural-sentences
- paper: http://papers.nips.cc/paper/5776-expressing-an-image-stream-with-a-sequence-of-natural-sentences.pdf
- paper: http://www.cs.cmu.edu/~gunhee/publish/nips15_stream2text.pdf
- author-page: http://www.cs.cmu.edu/~gunhee/
- github: https://github.com/cesc-park/CRCN

Multimodal Pivots for Image Caption Translation

- intro: ACL 2016
- arxiv: http://arxiv.org/abs/1601.03916

Image Captioning with Deep Bidirectional LSTMs

- intro: ACMMM 2016
- arxiv: http://arxiv.org/abs/1604.00790
- github(Caffe): https://github.com/deepsemantic/image_captioning
- demo: https://youtu.be/a0bh9_2LE24

Encode, Review, and Decode: Reviewer Module for Caption Generation

Review Network for Caption Generation

- intro: NIPS 2016
- arxiv: https://arxiv.org/abs/1605.07912
- github: https://github.com/kimiyoung/review_net

Attention Correctness in Neural Image Captioning

- arxiv: http://arxiv.org/abs/1605.09553

Image Caption Generation with Text-Conditional Semantic Attention

- arxiv: https://arxiv.org/abs/1606.04621
- github: https://github.com/LuoweiZhou/e2e-gLSTM-sc

DeepDiary: Automatic Caption Generation for Lifelogging Image Streams

- intro: ECCV International Workshop on Egocentric Perception, Interaction, and Computing
- arxiv: http://arxiv.org/abs/1608.03819

phi-LSTM: A Phrase-based Hierarchical LSTM Model for Image Captioning

- intro: ACCV 2016
- arxiv: http://arxiv.org/abs/1608.05813

Captioning Images with Diverse Objects

- arxiv: http://arxiv.org/abs/1606.07770

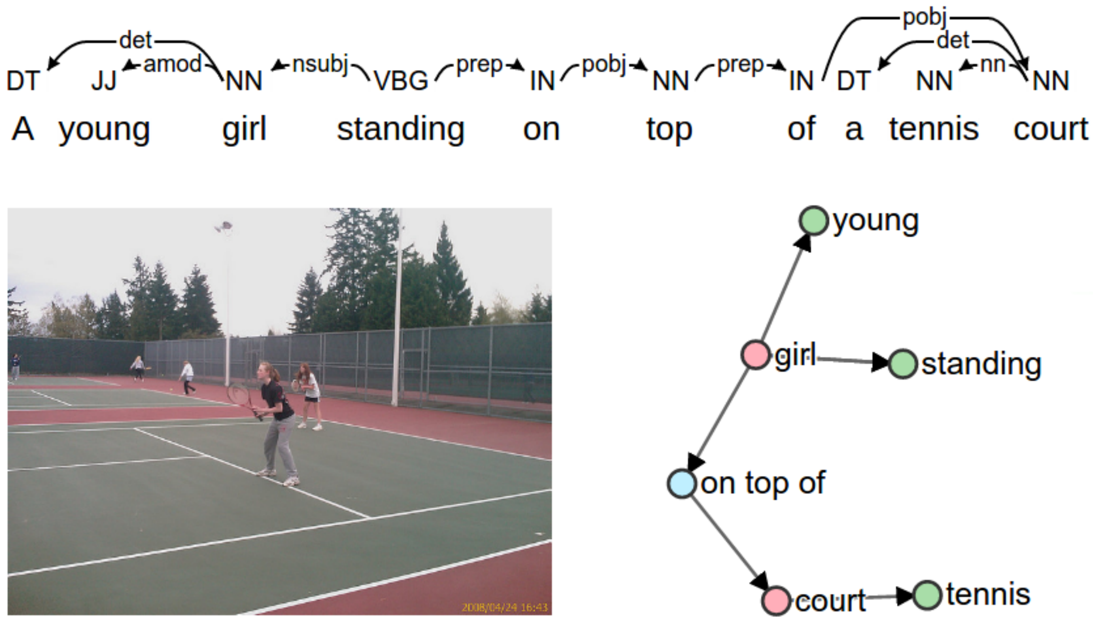
Learning to generalize to new compositions in image understanding

- arxiv: http://arxiv.org/abs/1608.07639

Generating captions without looking beyond objects

- intro: ECCV2016 2nd Workshop on Storytelling with Images and Videos (VisStory)
- arxiv: https://arxiv.org/abs/1610.03708

SPICE: Semantic Propositional Image Caption Evaluation



- intro: ECCV 2016
- project page: http://www.panderson.me/spice/
- paper: http://www.panderson.me/images/SPICE.pdf
- github: https://github.com/peteanderson80/SPICE

Boosting Image Captioning with Attributes

- arxiv: https://arxiv.org/abs/1611.01646

Bootstrap, Review, Decode: Using Out-of-Domain Textual Data to Improve Image Captioning

- arxiv: https://arxiv.org/abs/1611.05321

A Hierarchical Approach for Generating Descriptive Image Paragraphs

- intro: Stanford University
- arxiv: https://arxiv.org/abs/1611.06607

Dense Captioning with Joint Inference and Visual Context

- intro: Snap Inc.
- arxiv: https://arxiv.org/abs/1611.06949

Optimization of image description metrics using policy gradient methods

- intro: University of Oxford & Google
- arxiv: https://arxiv.org/abs/1612.00370

Areas of Attention for Image Captioning

- arxiv: https://arxiv.org/abs/1612.01033

Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning

- intro: CVPR 2017
- arxiv: https://arxiv.org/abs/1612.01887
- github: https://github.com/jiasenlu/AdaptiveAttention

Recurrent Image Captioner: Describing Images with Spatial-Invariant Transformation and Attention Filtering

- arxiv: https://arxiv.org/abs/1612.04949

Recurrent Highway Networks with Language CNN for Image Captioning

- arxiv: https://arxiv.org/abs/1612.07086

Top-down Visual Saliency Guided by Captions

- arxiv: https://arxiv.org/abs/1612.07360
- github: https://github.com/VisionLearningGroup/caption-guided-saliency

MAT: A Multimodal Attentive Translator for Image Captioning

https://arxiv.org/abs/1702.05658

Deep Reinforcement Learning-based Image Captioning with Embedding Reward

- intro: Snap Inc & Google Inc
- arxiv: https://arxiv.org/abs/1704.03899

Attend to You: Personalized Image Captioning with Context Sequence Memory Networks

- intro: CVPR 2017

- arxiv: https://arxiv.org/abs/1704.06485
- github: https://github.com/cesc-park/attend2u

Punny Captions: Witty Wordplay in Image Descriptions

https://arxiv.org/abs/1704.08224

Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner

https://arxiv.org/abs/1705.00930

Actor-Critic Sequence Training for Image Captioning

- intro: Queen Mary University of London & Yang’s Accounting Consultancy Ltd
- keywords: actor-critic reinforcement learning
- arxiv: https://arxiv.org/abs/1706.09601

What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?

- intro: Proceedings of the 10th International Conference on Natural Language Generation (INLG'17)
- arxiv: https://arxiv.org/abs/1708.02043

Stack-Captioning: Coarse-to-Fine Learning for Image Captioning

https://arxiv.org/abs/1709.03376

Self-Guiding Multimodal LSTM - when we do not have a perfect training dataset for image captioning

https://arxiv.org/abs/1709.05038

Object Descriptions

Generation and Comprehension of Unambiguous Object Descriptions

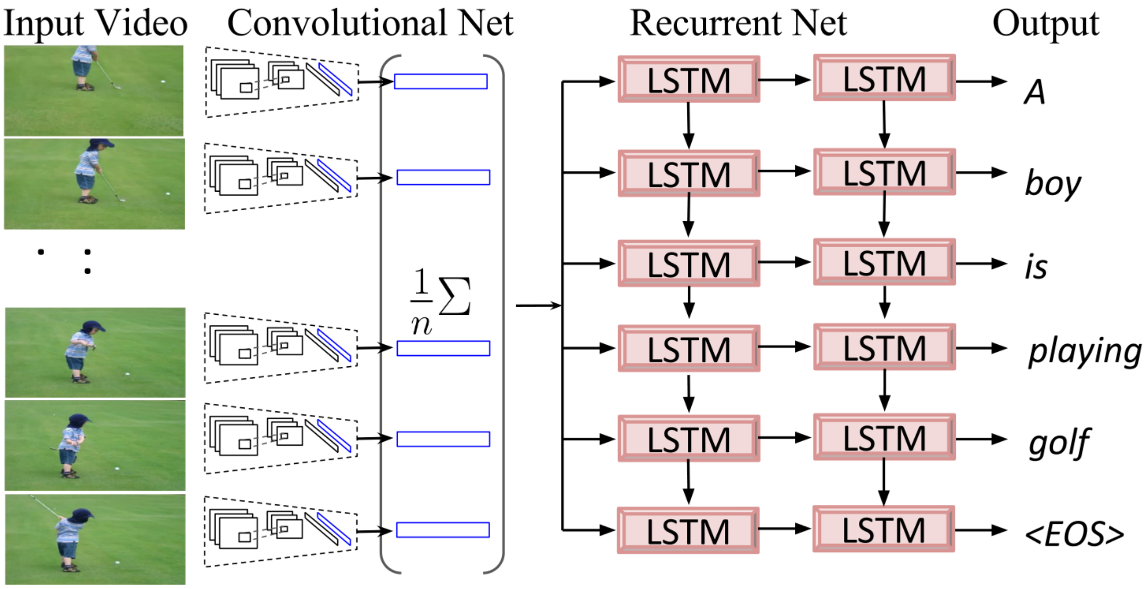
- arxiv: https://arxiv.org/abs/1511.02283
- github: https://github.com/mjhucla/Google_Refexp_toolbox

Video Captioning / Description

Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework

- intro: AAAI 2015
- paper: http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Pan_Jointly_Modeling_Embedding_CVPR_2016_paper.pdf
- paper: http://web.eecs.umich.edu/~jjcorso/pubs/xu_corso_AAAI2015_v2t.pdf

Translating Videos to Natural Language Using Deep Recurrent Neural Networks

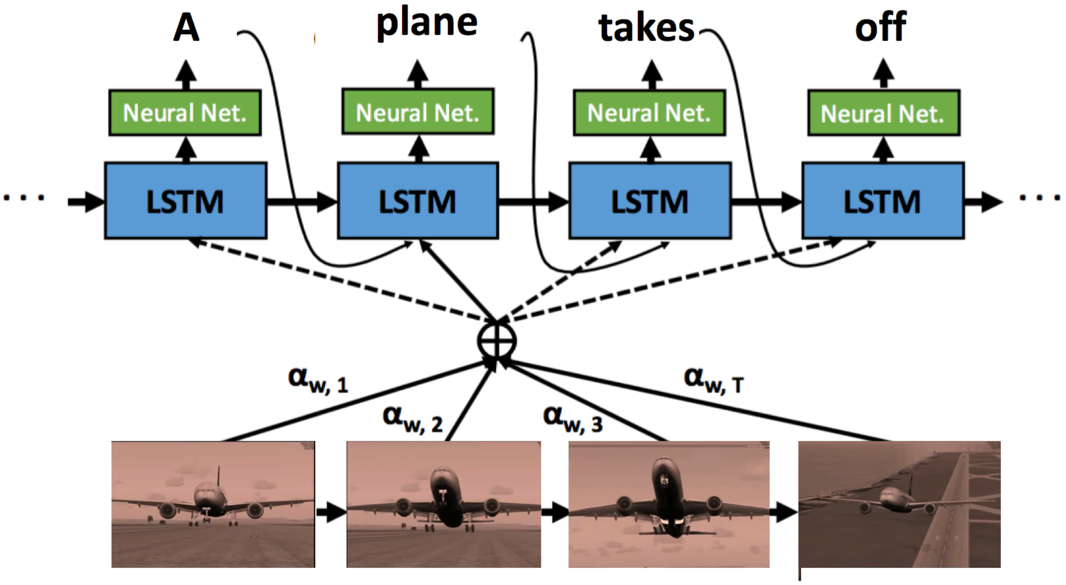


- intro: NAACL-HLT 2015 camera ready
- project page: https://www.cs.utexas.edu/~vsub/naacl15_project.html
- arxiv: http://arxiv.org/abs/1412.4729
- slides: https://www.cs.utexas.edu/~vsub/pdf/Translating_Videos_slides.pdf
- code+data: https://www.cs.utexas.edu/~vsub/naacl15_project.html#code

Describing Videos by Exploiting Temporal Structure

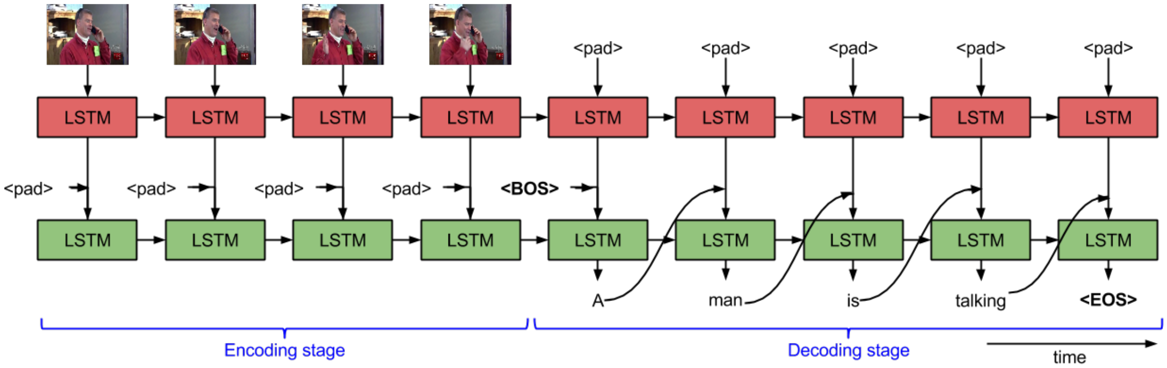
- arxiv: http://arxiv.org/abs/1502.08029
- github: https://github.com/yaoli/arctic-capgen-vid

SA-tensorflow: Soft attention mechanism for video caption generation



- github: https://github.com/tsenghungchen/SA-tensorflow

Sequence to Sequence -- Video to Text



- intro: ICCV 2015. S2VT
- project page: http://vsubhashini.github.io/s2vt.html
- arxiv: http://arxiv.org/abs/1505.00487
- slides: https://www.cs.utexas.edu/~vsub/pdf/S2VT_slides.pdf
- github(Caffe): https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt
- github(TensorFlow): https://github.com/jazzsaxmafia/video_to_sequence

Jointly Modeling Embedding and Translation to Bridge Video and Language

- arxiv: http://arxiv.org/abs/1505.01861

Video Description using Bidirectional Recurrent Neural Networks

- arxiv: http://arxiv.org/abs/1604.03390

Bidirectional Long-Short Term Memory for Video Description

- arxiv: https://arxiv.org/abs/1606.04631

3 Ways to Subtitle and Caption Your Videos Automatically Using Artificial Intelligence

- blog: http://photography.tutsplus.com/tutorials/3-ways-to-subtitle-and-caption-your-videos-automatically-using-artificial-intelligence--cms-26834

Frame- and Segment-Level Features and Candidate Pool Evaluation for Video Caption Generation

- arxiv: http://arxiv.org/abs/1608.04959

Grounding and Generation of Natural Language Descriptions for Images and Videos

- intro: Anna Rohrbach. Allen Institute for Artificial Intelligence (AI2)
- youtube: https://www.youtube.com/watch?v=fE3FX8FowiU

Video Captioning and Retrieval Models with Semantic Attention

- intro: Winner of three (fill-in-the-blank, multiple-choice test, and movie retrieval) out of four tasks of the LSMDC 2016 Challenge (Workshop in ECCV 2016)
- arxiv: https://arxiv.org/abs/1610.02947

Spatio-Temporal Attention Models for Grounded Video Captioning

- arxiv: https://arxiv.org/abs/1610.04997

Video and Language: Bridging Video and Language with Deep Learning

- intro: ECCV-MM 2016. captioning, commenting, alignment
- slides: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/10/Video-and-Language-ECCV-MM-2016-Tao-Mei-Pub.pdf

Recurrent Memory Addressing for describing videos

- arxiv: https://arxiv.org/abs/1611.06492

Video Captioning with Transferred Semantic Attributes

- arxiv: https://arxiv.org/abs/1611.07675

Adaptive Feature Abstraction for Translating Video to Language

- arxiv: https://arxiv.org/abs/1611.07837

Semantic Compositional Networks for Visual Captioning

- intro: CVPR 2017. Duke University & Tsinghua University & MSR
- arxiv: https://arxiv.org/abs/1611.08002
- github: https://github.com/zhegan27/SCN_for_video_captioning

Hierarchical Boundary-Aware Neural Encoder for Video Captioning

- arxiv: https://arxiv.org/abs/1611.09312

Attention-Based Multimodal Fusion for Video Description

- arxiv: https://arxiv.org/abs/1701.03126

Weakly Supervised Dense Video Captioning

- intro: CVPR 2017
- arxiv: https://arxiv.org/abs/1704.01502

Generating Descriptions with Grounded and Co-Referenced People

- intro: CVPR 2017. movie description
- arxiv: https://arxiv.org/abs/1704.01518

Multi-Task Video Captioning with Video and Entailment Generation

- intro: ACL 2017. UNC Chapel Hill
- arxiv: https://arxiv.org/abs/1704.07489

Dense-Captioning Events in Videos

- project page: http://cs.stanford.edu/people/ranjaykrishna/densevid/
- arxiv: https://arxiv.org/abs/1705.00754

Hierarchical LSTM with Adjusted Temporal Attention for Video Captioning

https://arxiv.org/abs/1706.01231

Reinforced Video Captioning with Entailment Rewards

- intro: EMNLP 2017. UNC Chapel Hill
- arxiv: https://arxiv.org/abs/1708.02300

End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering

- intro: CVPR 2017. Winner of three (fill-in-the-blank, multiple-choice test, and movie retrieval) out of four tasks of the LSMDC 2016 Challenge
- arxiv: https://arxiv.org/abs/1610.02947
- slides: https://drive.google.com/file/d/0B9nOObAFqKC9aHl2VWJVNFp1bFk/view

From Deterministic to Generative: Multi-Modal Stochastic RNNs for Video Captioning

https://arxiv.org/abs/1708.02478

Projects

Learning CNN-LSTM Architectures for Image Caption Generation: An implementation of CNN-LSTM image caption generator architecture that achieves close to state-of-the-art results on the MSCOCO dataset.

- github: https://github.com/mosessoh/CNN-LSTM-Caption-Generator

screengrab-caption: an openframeworks app that live-captions your desktop screen with a neural net

- intro: openframeworks app which grabs your desktop screen, then sends it to darknet for captioning. works great with video calls.
- github: https://github.com/genekogan/screengrab-caption

Tools

CaptionBot (Microsoft)

- website: https://www.captionbot.ai/

Blogs

Captioning Novel Objects in Images

http://bair.berkeley.edu/jacky/2017/08/08/novel-object-captioning/