State of the Art Presentation
Visual Storytelling

CS 698N: Recent Advances in Computer Vision

Vasu Sharma Nishant Rai Amlan Kar

[1]Department of Computer Science
Indian Institute of Technology, Kanpur

Instructor: Gaurav Sharma

Outline

The Problem : Introduction

- Introduced by Huang et al [1] from Microsoft Research at NAACL-2016
- Problem of mapping sequential images to sequential descriptive sentences
- Aim is to generate story like narrations

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

Figure: Visual Storytelling vs Caption generation

Types of Tasks

Image Sequence descriptions can be produced by a variety of approaches:

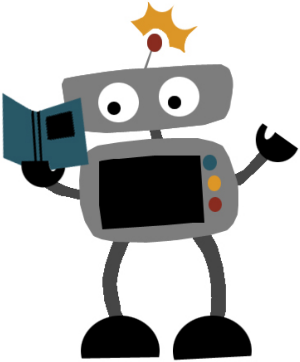
1. Descriptions of images in-isolation (**DII**)
2. Descriptions of images-in sequence (**DIS**)
3. Stories for images-in sequence (**SIS**)

DII	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
DIS	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Figure: Descriptions generated by DII, DIS and SIS approaches

Challenges

- Learning Human like narrative language
- Ability to remember long term context from images and be able to connect their ideas together
- Only Jamie Kiros' Neural Storyteller[2] comes close to achieving this using their SkipThought vectors[3]



Dataset Collection and Description

- **81,743** unique photos in **20,211** sequences with captions and narrative sequences
- Flickr API used to extract photo albums
- Amazon Mechanical Turkers used to get narrative stories and isolated captions
- Data Post-processing performed

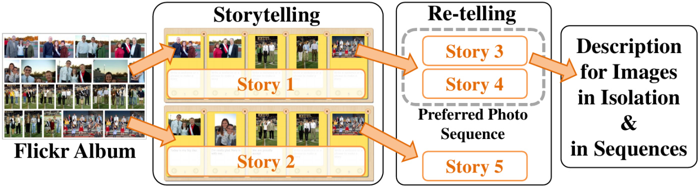


Figure: Dataset Collection Crowdsourcing Workflow

Evaluation Metrics

- Evaluating the quality of the generated stories is a non trivial task.
- Intuitive way involves **comparison** with good (**human-made**) model stories. But manual evaluation not possible for large sets.
- Need of automatic methods for such evaluations.
- Popular metrics which assign a **score** to the candidate (based on human-made ground truths) include **BLEU**, **METEOR**.

BLEU [4]

- Account for adequacy by calculating word-match precision, account for fluency by computing n-gram precisions
- Smaller sentences get higher scores, thus a length based penalty introduced to prevent it
- More reference human samples result in better and accurate scores
- Designed to approximate human judgement at a corpus level, and performs badly if used to evaluate the quality of individual sentences
- Example:
  - "There is a cat on the mat; The cat is on the mat" vs "the the the the the the"
  - "There is a cat on the mat; The cat is on the mat" vs "the cat"

METEOR [6]

- Consistently outperforms BLEU in correlation with human judgments
- Sentence alignment takes variability into account via stemming and synonymy matching
- Combine Recall and Precision as weighted score components
- Align candidate with each reference and take score of the best pairing
- Consider the fragmentation of the candidate-reference alignment

https://www.sharelatex.com/project/57bf30d8ead3386f0bacb570

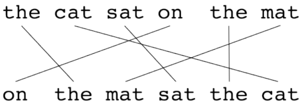


Figure: Image from Wikipedia [5]

Basic Approach

- A sequence-to-sequence recurrent neural net (**seq2seq**)[7] used for story generation
- Image sequence **encoded** by running an RNN over image representations (e.g. the activations of another pre-trained model). Used as the **initial** hidden state to the story decoder model
- The story decoder model produces the story **one word** at a time from the training data vocabulary
- **GRUs** are used as the image encoders and story decoders

Heuristics Used

- **METEOR** score used for comparing model performance
- Multiple heuristics used to further improve results including,
  - Lower **Beam Search** size
  - Avoid **duplicates**
  - Penalize **Visually-Grounded** words

Results

As discussed earlier, METEOR metric used for evaluation

Beam=10	Greedy	-Dups	+Grounded
23.55	19.10	19.21	–

Figure: Scores for generated captions per-image

Beam=10	Greedy	-Dups	+Grounded
23.13	27.76	30.11	31.42

Figure: Scores for generated stories

Bibliography I

T. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell.
Visual storytelling, 2016.
Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Neural story teller.
`https://github.com/ryankiros/neural-storyteller`.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler.
Skip-thought vectors.
In *Advances in neural information processing systems*, pages 3294–3302, 2015.

Bibliography II

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Meteor wiki.
https://en.wikipedia.org/wiki/METEOR.

Satanjeev Banerjee and Alon Lavie.
Meteor: An automatic metric for mt evaluation with improved correlation with human judgments.
2005.

Bibliography III

Ilya Sutskever, Oriol Vinyals, and Quoc V Le.
Sequence to sequence learning with neural networks.
In *Advances in neural information processing systems*, pages 3104–3112, 2014.