Final Project 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

1. The Problem
   - Introduction
   - Types of Tasks
2. Our Approach
3. Training and Implementation details
4. Results
5. Story generation example
6. Summary of contributions

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

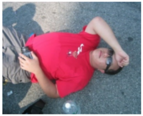
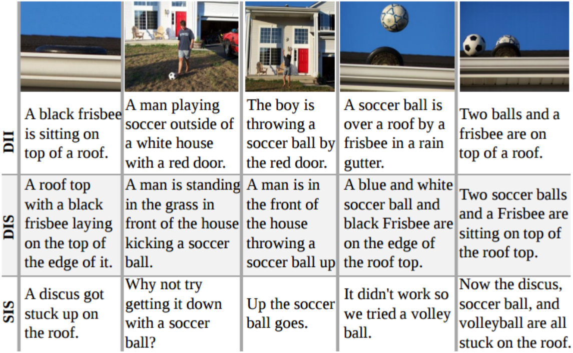
			
DII	A group of people that are sitting next to each other.	Adult male wearing sunglasses lying down on black pavement.	The sun is setting over the ocean and mountains.
SIS	Having a good time bonding and talking.	[M] got exhausted by the heat.	Sky illuminated with a brilliance of gold and orange hues.

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

Our Approach

- Data pre-processing and creating of TensorFlow data pipeline
- Image embedding using inception network
- GRU based encoding of image sequence run over the sequence in reverse order. Used as initial stae for decoder
- Decoding the image encodings using a GRU decoder and beam search to produce stories word by word
- Evaluation of generated stories using the METEOR metric

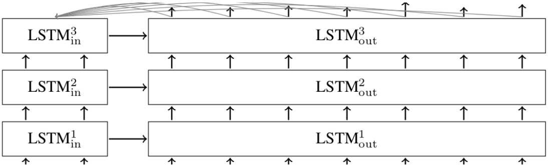

$\mathrm{LSTM}^3_{\mathrm{in}}$
$\mathrm{LSTM}^3_{\mathrm{out}}$
$\mathrm{LSTM}^2_{\mathrm{in}}$
$\mathrm{LSTM}^2_{\mathrm{out}}$
$\mathrm{LSTM}^1_{\mathrm{in}}$
$\mathrm{LSTM}^1_{\mathrm{out}}$

Figure: Multi layer LSTM story generation Architecture

Training and Implementation details

- The dataset contains **81,743** unique photos in **20,211** sequences with captions and narrative sequences
- Inception network pre-trained on MSCOCO and produces image embeddings
- Inception network initially frozen and encoder-decoder trained
- Inception network unfrozen and whole network is finetuned together for the Visual Storytelling task
- All encoder states used for decoding
- Bidirectional GRU's used for decoder network
- Custom built beam search to support multi threading
- Heuristics like preventing duplication sentence generation and variable beam width in beam search used

Results

Beam Beam Width	No repeat Heuristic	METEOR	Bleu	CIDEr	ROUGUE_L
1	Yes	0.071	0.173	0.060	0.153
2	Yes	0.082	0.209	0.097	0.160
3	Yes	0.084	0.217	0.094	0.163
4	Yes	0.083	0.217	0.089	0.163
5	Yes	0.082	0.214	0.092	0.161
1	No	0.062	0.152	0.036	0.148
3	No	0.063	0.159	0.038	0.146
5	No	0.060	0.152	0.030	0.144

Table: Results on the Visual Storytelling task using our approach

Story generation example

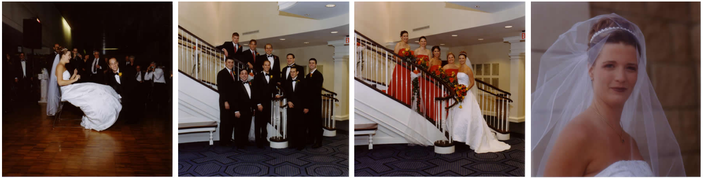


Figure: Example Image sequence

Our caption: *the bride and groom were very happy to be getting married . the family is having a great time . the couple was excited to see their new friends . The bride and her bridesmaids looked absolutely gorgeous . the bride was happy to be there.*

Story generation example

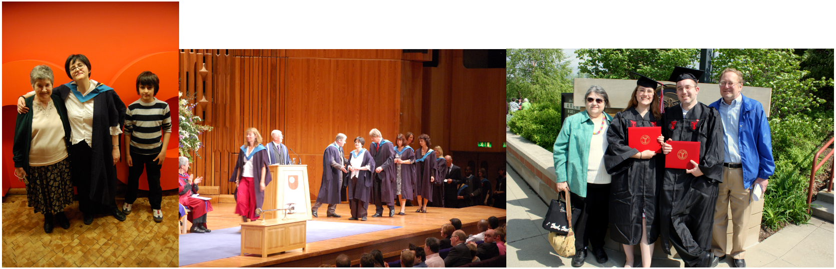


Figure: Example Image sequence

Our caption: *the family was very nervous. the students were excited to be graduating . the graduation ceremony was held and everyone was very happy . the graduates were very proud of their accomplishments . the group of friends posed for a picture .*

Summary of contributions

Feature	Proposed	Mid sem progress	Final Progress
Using seq2seq model for generating captions (Implemented from scratch)	✓	✓	✓
Producing descriptions for images in sequence	✓	✓	✓
Replicating the State of the art paper results	✓	✗	✓
Decoding using all encoder states	✓	✗	✓
Bi-directional connections in LSTM	✓	✗	✓
Custom Implementation of Beam Search	✓	✗	✓

Table: Comparison table between proposal and mid term and final progress reports

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, M. Mitchell
  Visual Storytelling.
  North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016

Thank You

Thank You!!! :)

Questions??