VIDEO DESCRIPTION

MADHAV CHAVDA (1401006)

PRERAK RAJA (1401041)

RATNESH SHAH(1401110)

VARAD BHOGAYATA(1401042)

OBJECTIVE

A monkey is pulling a dog's tail and is chased by the dog.

- Given an input video, generate the crux of the entire video

MOTIVATION

robotic vision

assist for blinded

incident report for surveillance

nultimedia search

movie description for blinded

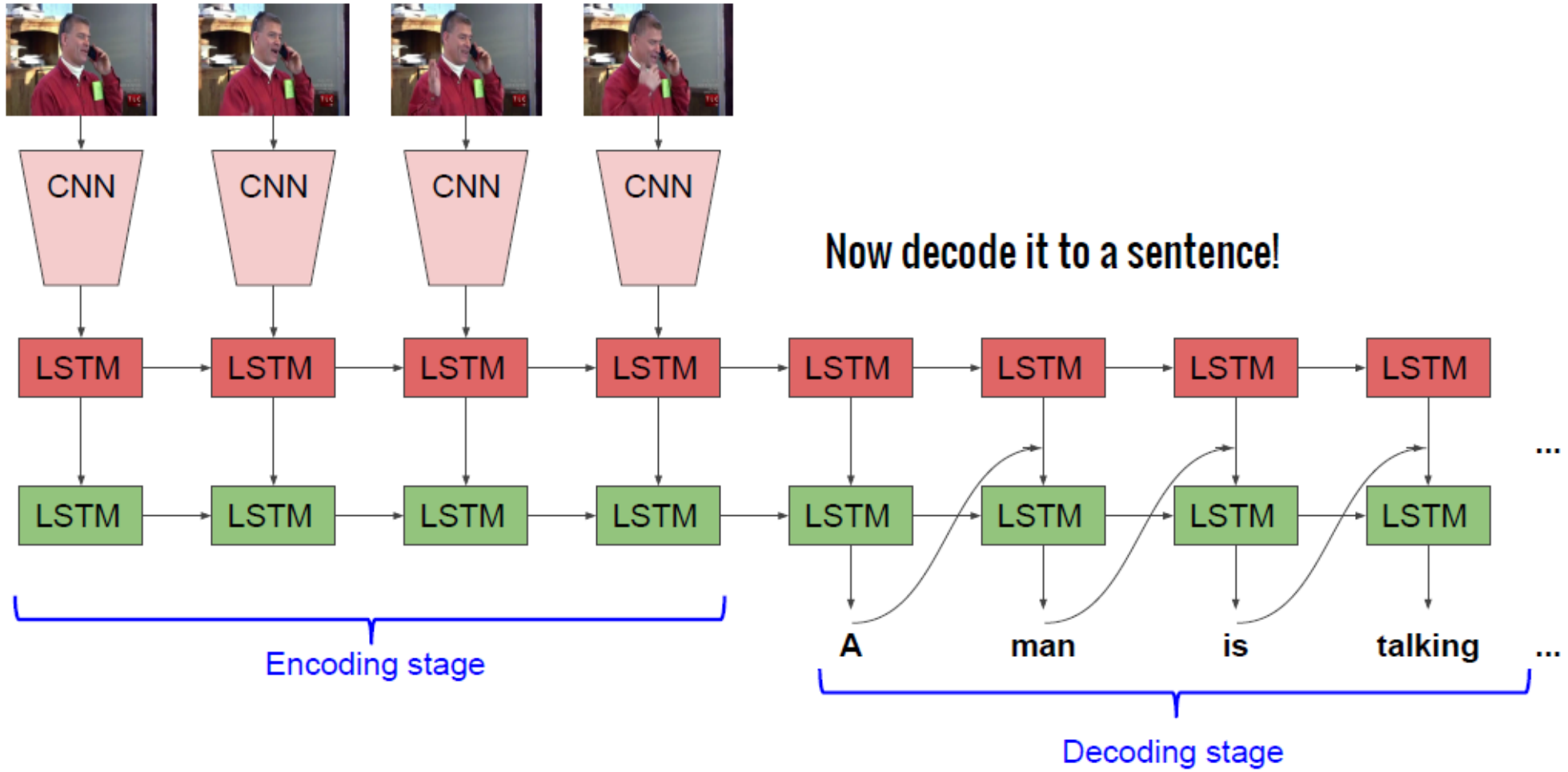
seeing chat bot

DATA-SET

MSVD (The Microsoft Video description corpus):

- The Microsoft Video description corpus, is a collection of Youtube clips collected on Mechanical Turk by requesting workers to pick short clips depicting a single activity.
- The videos were then used to elicit single sentence descriptions from annotators.
- About 120K sentences (including different languages)
- Around 2000 videos
- Example of video from the dataset

LSTMS FOR SEQUENCE MODELING



APPROACH

Preprocessing

- Resizing video to same dimension (224 x 224 x3)
- Subsampling (50 frames for each video)

Extraction of feature vector for each frame

- VGG16 to extract features
- Features from fc7 layer stored in numpy file for each video (50 frames: 50x4096)

Two stacked LSTM (Seq 2 Seq)

- Learn a representation of a sequence of frames in order to decode it into a sentence that describes the event in the video. The top LSTM layer models visual feature inputs.
- The second LSTM layer models language given the text input and the hidden representation of the video sequence.

Sequence to Sequence Model

Train

- **Getting video data:**
  - Train: 90% , Test: 10%
  - From Video Corpus file, choose only English language
  - Separate out Video-ID, Video-Caption
- **Build Vocabulary:**
  - Generate two numpy files : WordToIndex and IndexToWord (if count of word is greater than predefined threshold, then store it in vocab)
  - Generated Vocab
- **Model Initialization:**
  - Initialize different parameters: image dimension, no. of hidden layers, no of words, batch size, no of lstm steps.
- **Build Model:**
  - Phase-1 : Read frames
  - Phase-2: Generate Captions
  - Calculate Loss
- **Training the model**

Test

- **Getting video data:**
  - Take test Video data
- **Load IndexToWord File**
- **Model Initialization:**
  - Initialize different parameters: image dimension, no. of hidden layers, no of words, batch size, no of lstm steps.
- **Generate Caption:**
  - Phase-1 : Read frames
  - Phase-2: Generate Captions (select word with maximum probability)

RESULTS

- project\YouTubeClips\jPBxl9gFqNY_110_117.avi
  - Ground truth: A man is pouring oil into a frying pan.
  - At 100th epoch: man is cooking
  - At 500th epoch: man is pouring some
  - At 900th epoch: man is putting oil into a
- project\YouTubeClips\X6uJyuD_Zso_3_17.avi
  - Ground truth: A man is chopping an onion.
  - At 100th epoch: person is slicing a
  - At 500th epoch: man is slicing
  - At 900th epoch: man is slicing

RESULTS

- project\YouTubeClips\229NvV0SRHw_0_5.avi
  - Ground truth: A baby is drinking from a cup
  - At 100th epoch: baby is eating a
  - At 500th epoch: young girl is playing with a
  - At 900th epoch: baby is eating
- project\YouTubeClips\Puh1n8DTKw8_2_9.avi
  - Ground truth: A baby girl is walking
  - At 100th epoch: baby is
  - At 500th epoch: baby is
  - At 900th epoch: baby is

DIFFICULTIES FACED

- In generating features we needed to extract 80 frames but since the dataset was large all the 80 frames were not able to load in memory. So we decided to extract 50 frames as the system was able to load it.
- In the sentence which are generated we can see that it is not complete. We think that this problem might be because padding or because we are using less no. of frames.
- Training Time required for 1000 epochs was found to be around 5 hours.
- Several inaccuracies in the generated sentence.
- When we generate the vocabulary, the threshold kept is 10. Now, if any name appears less than the threshold we may lose some keywords.

EVALUATION CRITERIA

- Appropriate Length
- Fidelity
- Salience
- Grammaticality
- Non-redundancy
- Structure and Coherence

METEOR

- $Precision\ P = \frac{no\ of\ words\ in\ ground\ truth\ \cap no\ of\ words\ in\ predicted\ sentence}{no\ of\ words\ in\ prediceted\ sentence}$
- $Recall\ R = \frac{no\ of\ words\ in\ ground\ truth\ \cap no\ of\ words\ in\ predicted\ sentence}{no\ of\ words\ in\ Ground\ Truth}$
- $F_{mean} = \frac{10PR}{R+9P}$
- $Fragmentation = chunks/matches$
- $Penalty\ p = 0.5(fragmentation)^3$
- $Score\ M = F_{mean} * (1-p)$

Example 1: Reference : the cat sat on the mat

Hypothesis: on the mat sat the cat

Score = 0.5

Example 2: Reference : the cat sat on the mat

Hypothesis: the cat was sat on the mat

Score = 0.964

ROUGE-L

- $Precision\ P = \frac{no\ of\ words\ in\ ground\ truth\ \cap no\ of\ words\ in\ predicted\ sentence}{no\ of\ words\ in\ prediceted\ sentence}$
- $Recall\ R = \frac{no\ of\ words\ in\ ground\ truth\ \cap no\ of\ words\ in\ predicted\ sentence}{no\ of\ words\ in\ Ground\ Truth}$
- $Score\ = \frac{(1+(1.2)^2)PR}{R+(1.2)^2P}$

Example 1: Reference : the cat sat on the mat

Hypothesis: on the mat sat the cat

Score = 0.5

Example 2: Reference : the cat sat on the mat

Hypothesis: the cat was sat on the mat

Score = 0.9360

EVALUATION (ROUGE-L)

Video \ Epochs	100th Epoch	500th Epoch	900th Epoch
Video-1	30.57	43.16	64.34
Video-2	19.30	41.92	41.92
Video-3	51.98	30.34	37.30
Video-4	53.40	53.40	53.40

MODEL EVALUATION RESULTS

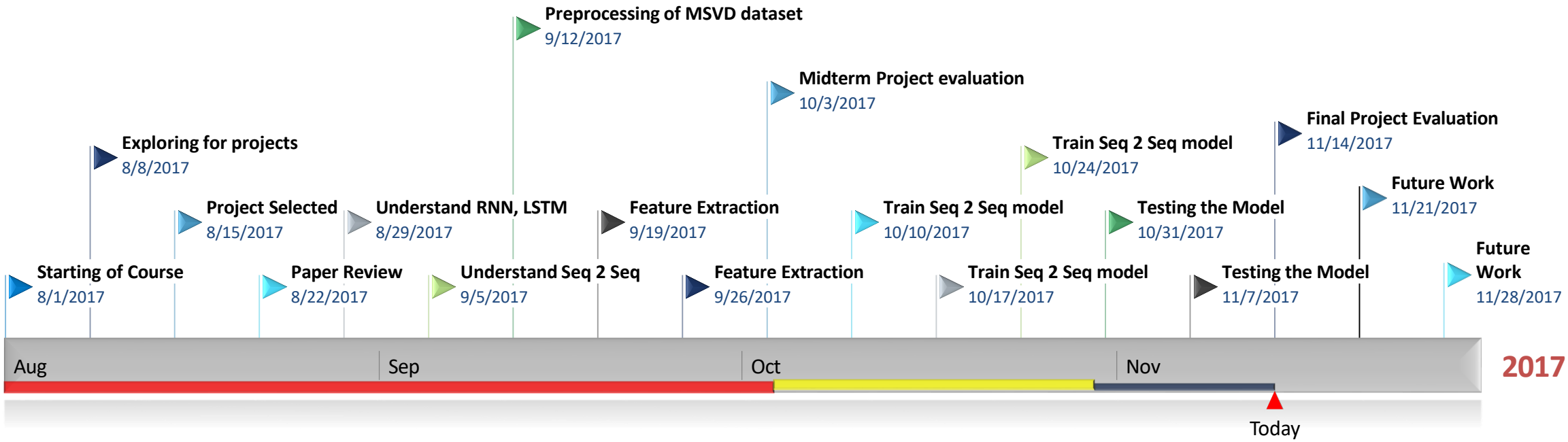
Method	Score
ROUGE-L	0.659
BLEU	0.55
METEOR	0.276
CIDEr	0.461

ROUGE- Recall-Oriented Understudy for Gisting Evaluation
BLEU- Bilingual Evaluation Understudy
METEOR -Metric for Evaluation of Translation with Explicit ORdering
CIDEr- Consensus-based Image Description Evaluation

LEARNING

- VGG-16 Model
- Recurrent Neural Network
- LSTM
- Image Captioning
- Sequence 2 Sequence Model
- METEOR Metric

TIMELINE


Starting of Course
8/1/2017
Exploring for projects
8/8/2017
Project Selected
8/15/2017
Paper Review
8/22/2017
Understand RNN, LSTM
8/29/2017
Understand Seq 2 Seq
9/5/2017
Preprocessing of MSVD dataset
9/12/2017
Feature Extraction
9/19/2017
Feature Extraction
9/26/2017
Midterm Project evaluation
10/3/2017
Train Seq 2 Seq model
10/10/2017
Train Seq 2 Seq model
10/17/2017
Train Seq 2 Seq model
10/24/2017
Testing the Model
10/31/2017
Testing the Model
11/7/2017
Final Project Evaluation
11/14/2017
Future Work
11/21/2017
Future Work
11/28/2017
Aug
Sep
Oct
Nov
2017
Today

REFERENCE

- arXiv:1505.00487 [cs.CV]
- https://www.youtube.com/watch?v=iX5V1WpxxkY
- https://vsubhashini.github.io/s2vt.html
- http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture10.pdf
- https://medium.com/towards-data-science/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d
- https://www.tensorflow.org/tutorials/seq2seq

THANK YOU