A11: Deep Video Understanding

Date of Presentation : May 6, 2022

VIDEO CAPTION GENERATION

Sunjana Ramana
UNI - sc4921
Department of Electrical Engineering
Columbia University
sc4921@columbia.edu

A boy is sitting on a car seat of a car

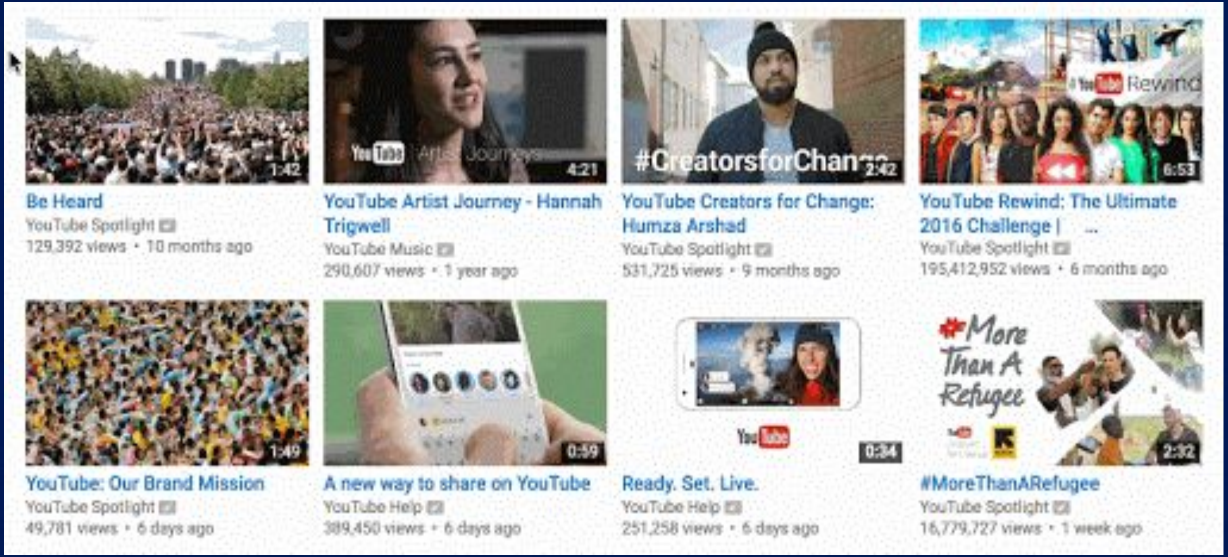
CONTENTS

Final Project

1. Motivation
2. Data
3. System Overview
4. Methodology
   i) Video to frames
   ii) Feature Extraction
   iii) Caption Preprocess
   iv) VTT Model
   v) Testing
5. Results and Analysis

MOTIVATION

> VIDEOS ARE EVERYWHERE
> VIDEOS ARE A MODE OF COMMUNICATION
> AI FOR VIDEOS IS ESSENTIAL



Be Heard
YouTube Spotlight
129,392 views · 10 months ago

YouTube Artist Journey - Hannah Trigwell
YouTube Music
290,607 views · 1 year ago

YouTube Creators for Change: Humza Arshad
YouTube Spotlight
531,725 views · 9 months ago

YouTube Rewind: The Ultimate 2016 Challenge | ...
YouTube Spotlight
195,412,952 views · 6 months ago

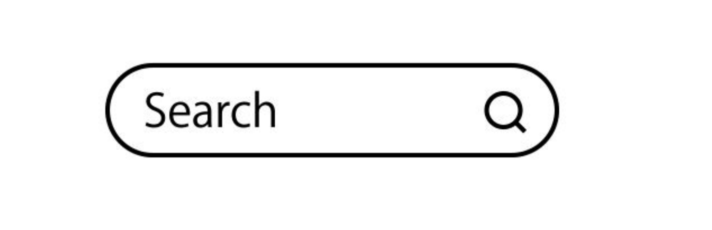
YouTube: Our Brand Mission
YouTube Spotlight
49,781 views · 6 days ago

A new way to share on YouTube
YouTube Help
389,450 views · 6 days ago

Ready. Set. Live.
YouTube Help
251,258 views · 6 days ago

#MoreThanARefugee
YouTube Spotlight
16,779,727 views · 1 week ago

VIDEO —----> TEXT



> EFFICIENT SEARCH ALGORITHMS

Traditional : Search by Tags/ Title

With VTT :

- Search by Videos
- Search by context
- Search by Content

> BETTER RECOMMENDATIONS FOR USERS

- Use VTT captions along with Tags/Title to build better recommendation sys

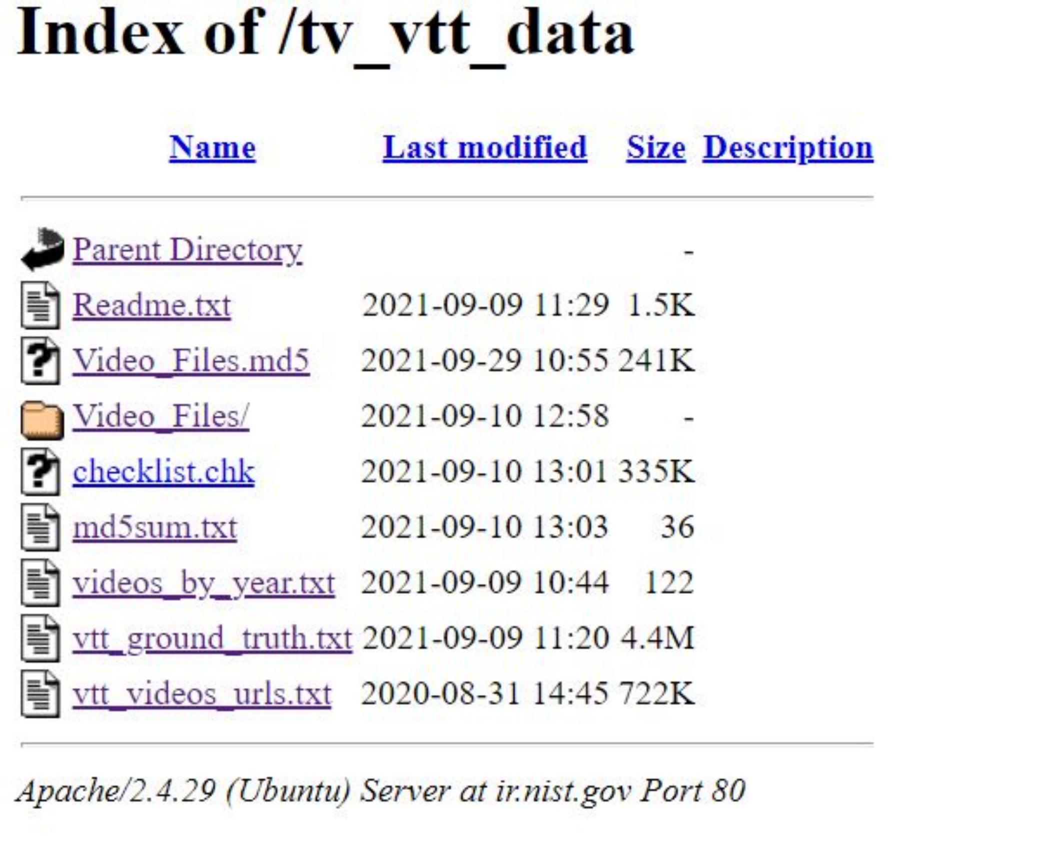
DATA SOURCE

> TRECVID

- Short Videos (3 to 10 seconds)
- TRECVID VTT task from 2016 to 2021
- Total of ~10000 videos
- There are 6478 URLs from Twitter Vine
- The remaining from Flicker and Vimeo
- Each video has between 2 and 5 captions

> 3 Vs of Big Data

Volume : ~10k videos
Velocity : Different Lengths
Variety : 3 different sources



Index of /tv_vtt_data

Name	Last modified	Size	Description
Parent Directory		-	
Readme.txt	2021-09-09 11:29	1.5K	
Video_Files.md5	2021-09-29 10:55	241K	
Video_Files/	2021-09-10 12:58	-	
checklist.chk	2021-09-10 13:01	335K	
md5sum.txt	2021-09-10 13:03	36	
videos_by_year.txt	2021-09-09 10:44	122	
vtt_ground_truth.txt	2021-09-09 11:20	4.4M	
vtt_videos_urls.txt	2020-08-31 14:45	722K	

Apache/2.4.29 (Ubuntu) Server at ir.nist.gov Port 80

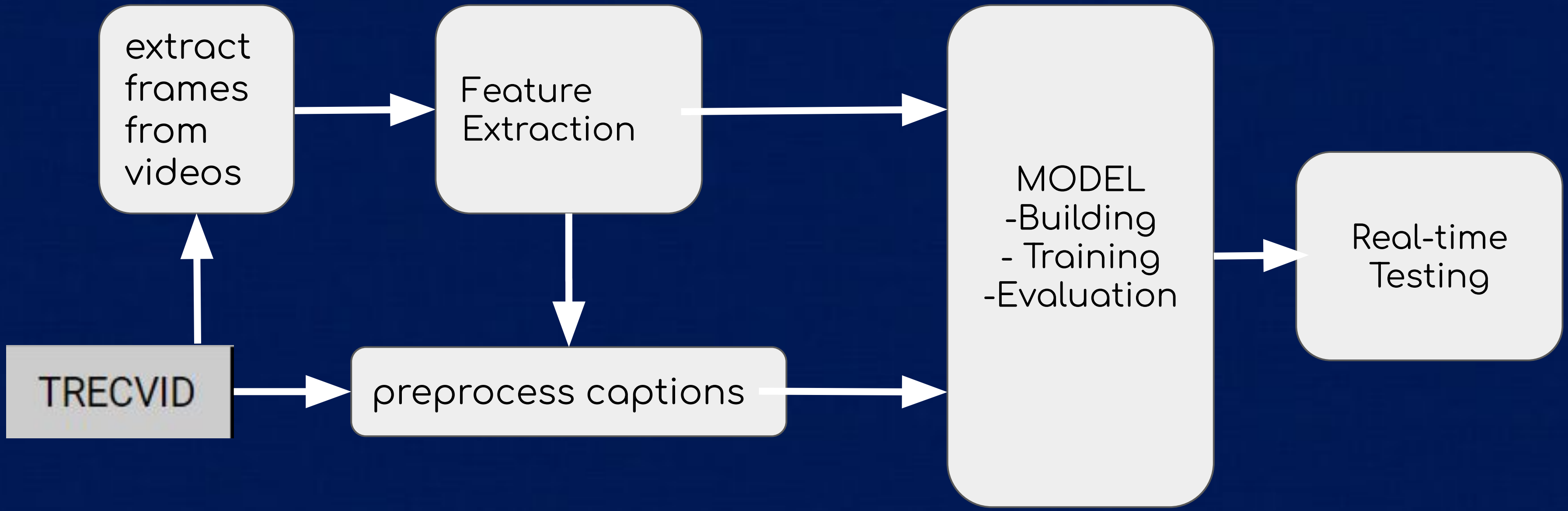
TRAIN DATA

vtt_videos_urls.txt

TEST DATA

Video_Files/xxx.webmd

SYSTEM OVERVIEW



extract frames from videos

Feature Extraction

MODEL
-Building
- Training
-Evaluation

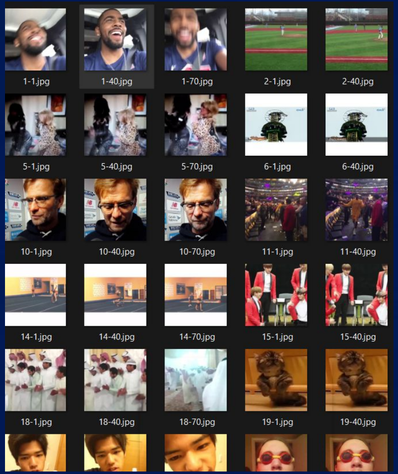
Real-time Testing

TRECVID

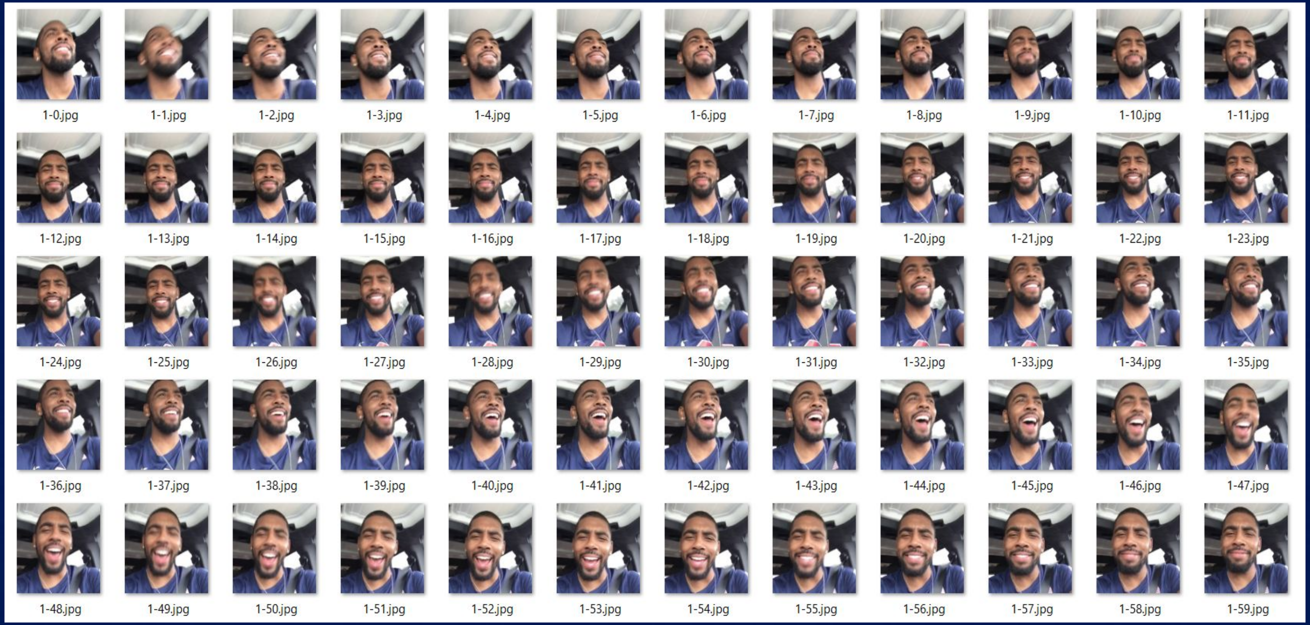
preprocess captions

VIDEO FRAME EXTRACTION : Video Lengths differ

BEFORE



NOW



VIDEO FRAME EXTRACTION :

> The training data consists of 6478 urls

- Length of each video in the url is different
- A constant number of frames was necessary
- Generally there are 24 frames in a second
- To accommodate 3 second videos about 80 frames are considered

> Now each video has 80 frames

Ex: ID1 —- 80 images
ID2 — 80 images
ID3 — 80 images

FEATURE EXTRACTION

> Used Pre-trained Models for Feature Extraction

    > Using VGG16

        - Features were stored as (80 , 4096) numpy arrays for each video

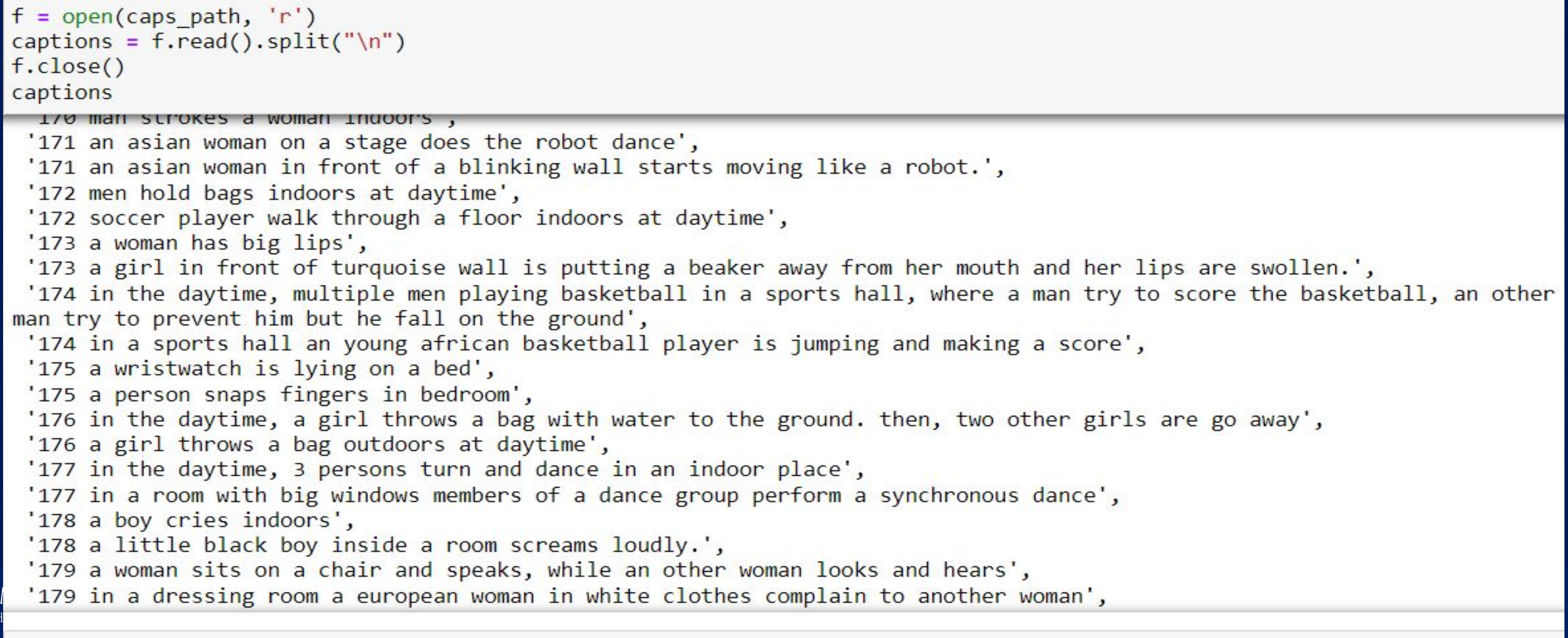
    > Using InceptionResnetV2

        - Features were stores as (80, 1536) numpy arrays

> Drawbacks : Time Consuming – 6478*(80,4096))

CAPTIONS CLEANING AND PREPROCESSING

> Maximum length - 37 , Minimum length - 3



```
f = open(caps_path, 'r')
captions = f.read().split("\n")
f.close()
captions
```

```
 '170 man strokes a woman indoors',
 '171 an asian woman on a stage does the robot dance',
 '171 an asian woman in front of a blinking wall starts moving like a robot.',
 '172 men hold bags indoors at daytime',
 '172 soccer player walk through a floor indoors at daytime',
 '173 a woman has big lips',
 '173 a girl in front of turquoise wall is putting a beaker away from her mouth and her lips are swollen.',
 '174 in the daytime, multiple men playing basketball in a sports hall, where a man try to score the basketball, an other
man try to prevent him but he fall on the ground',
 '174 in a sports hall an young african basketball player is jumping and making a score',
 '175 a wristwatch is lying on a bed',
 '175 a person snaps fingers in bedroom',
 '176 in the daytime, a girl throws a bag with water to the ground. then, two other girls are go away',
 '176 a girl throws a bag outdoors at daytime',
 '177 in the daytime, 3 persons turn and dance in an indoor place',
 '177 in a room with big windows members of a dance group perform a synchronous dance',
 '178 a boy cries indoors',
 '178 a little black boy inside a room screams loudly.',
 '179 a woman sits on a chair and speaks, while an other woman looks and hears',
 '179 in a dressing room a european woman in white clothes complain to another woman',
```

CAPTIONS CLEANING AND PREPROCESSING

> Problem of excessive padding

- This would lead to the model predicting blank spaces
- Therefore lengths between 6-10 were selected and ‘ < start > ‘ and ‘<end> ‘ tags were added to the captions

> for 6478 urls :

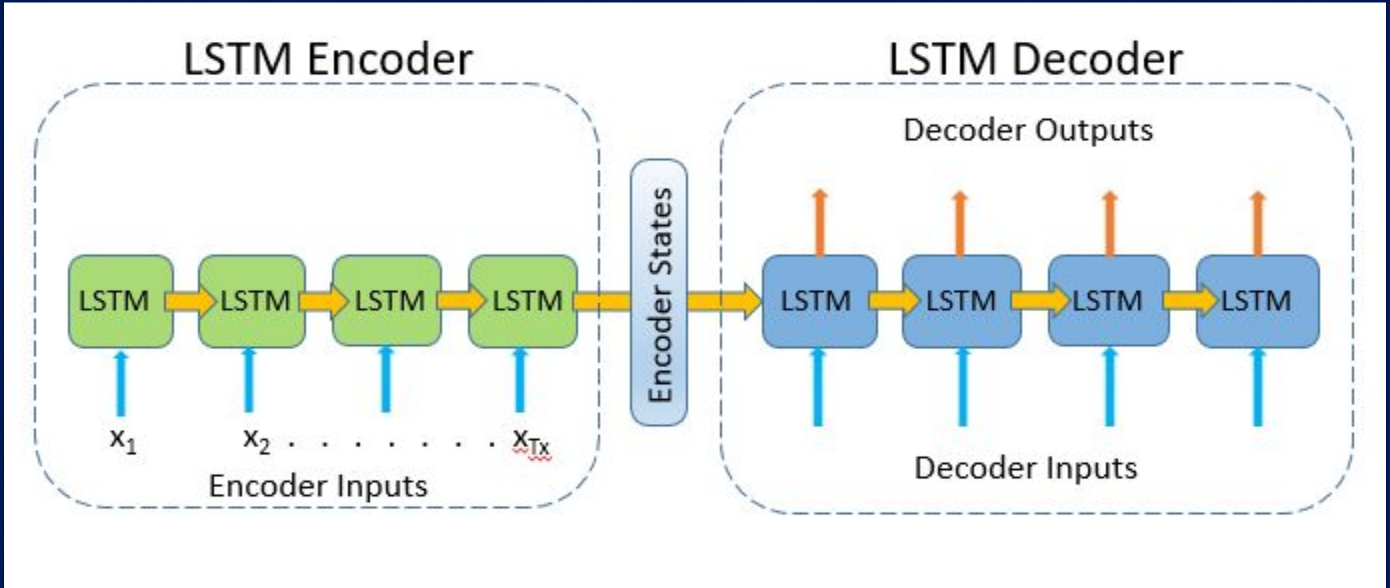
url1 — 80 frames paired with 2-5 captions

url2 — 80 frames each paired with 2-5 captions

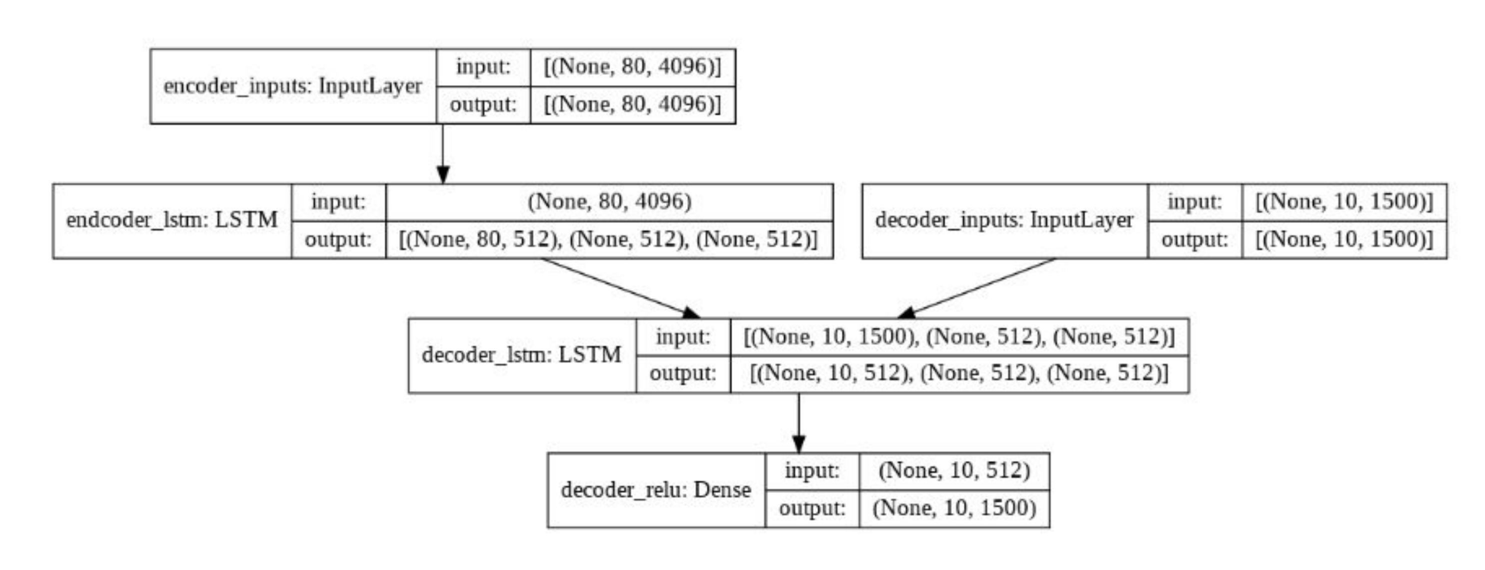
> train_val split = 0.85

MODEL for TRAINING :Encoder Decoder

>



MODEL for TRAINING :Encoder Decoder



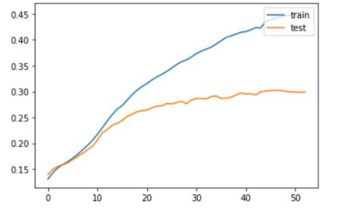
encoder_inputs: InputLayer | input: [(None, 80, 4096)] | output: [(None, 80, 4096)]

endcoder_lstm: LSTM | input: (None, 80, 4096) | output: [(None, 80, 512), (None, 512), (None, 512)]

decoder_inputs: InputLayer | input: [(None, 10, 1500)] | output: [(None, 10, 1500)]

decoder_lstm: LSTM | input: [(None, 10, 1500), (None, 512), (None, 512)] | output: [(None, 10, 512), (None, 512), (None, 512)]

decoder_relu: Dense | input: (None, 10, 512) | output: (None, 10, 1500)

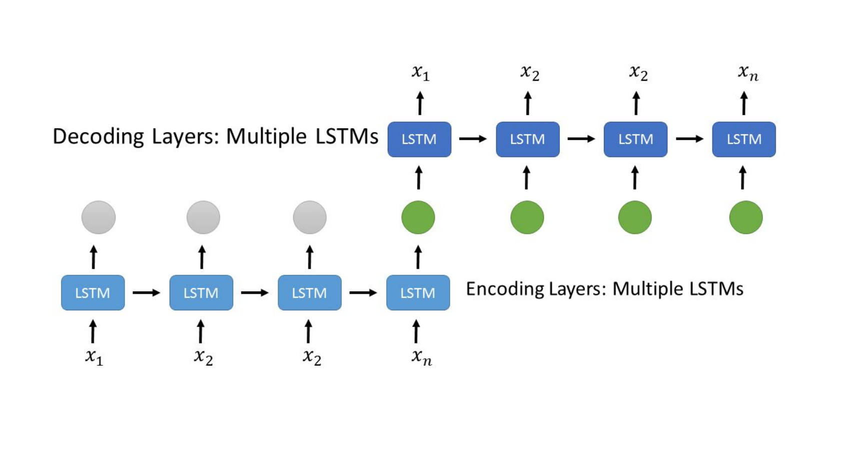


TESTING:

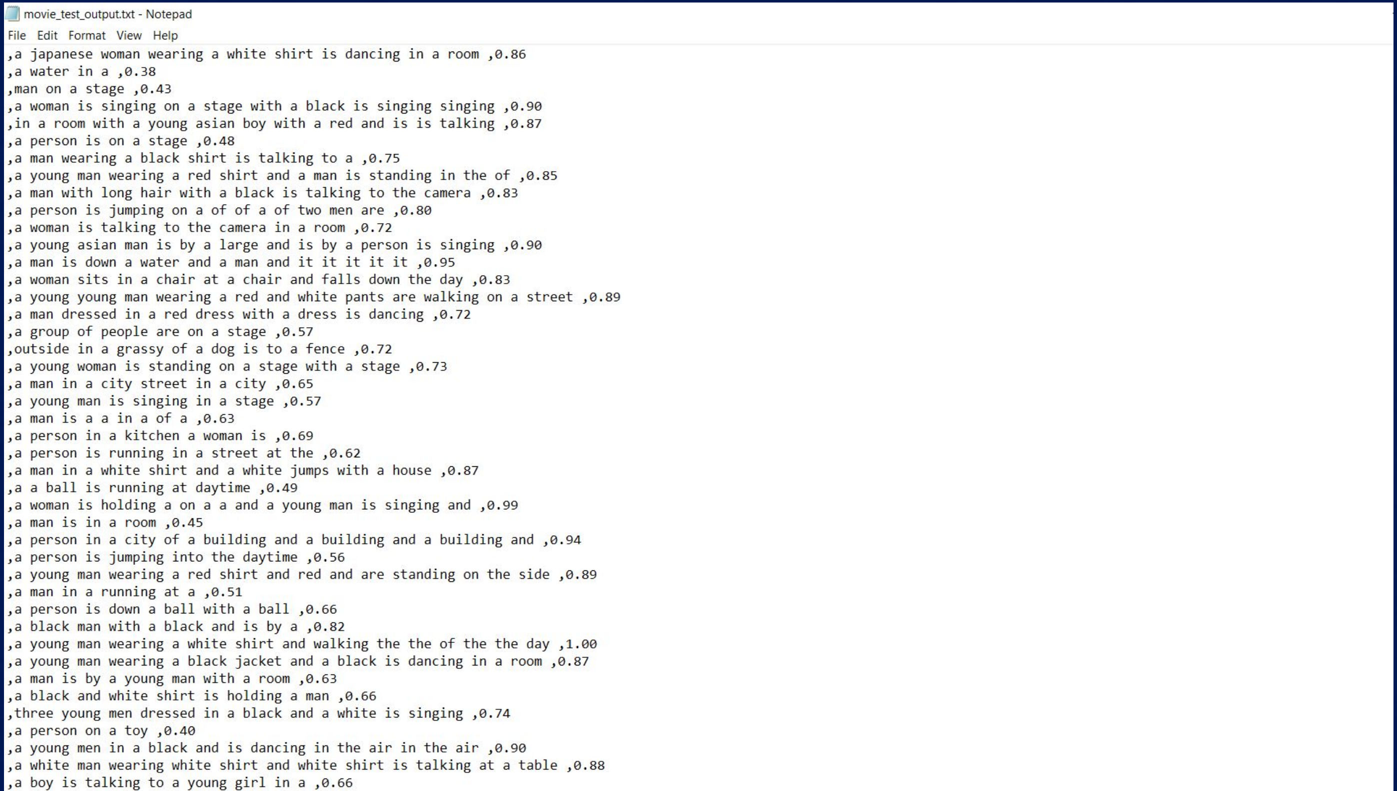
> Combination of Greedy Search and Beam Search Decoding

greedy decoding calculate the best option based on the very next word/token only

beam search checks for multiple word/tokens into the future and assesses the quality of all of these tokens combined.

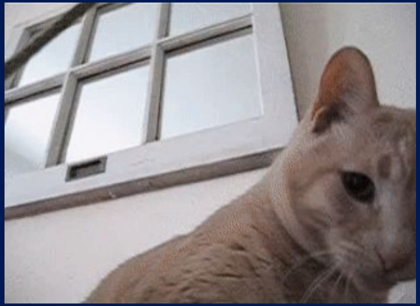


RESULTS:

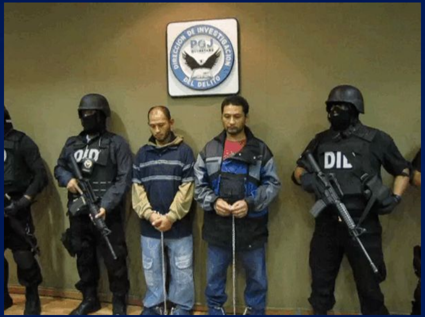


```
,a japanese woman wearing a white shirt is dancing in a room ,0.86
,a water in a ,0.38
,man on a stage ,0.43
,a woman is singing on a stage with a black is singing singing ,0.90
,in a room with a young asian boy with a red and is is talking ,0.87
,a person is on a stage ,0.48
,a man wearing a black shirt is talking to a ,0.75
,a young man wearing a red shirt and a man is standing in the of ,0.85
,a man with long hair with a black is talking to the camera ,0.83
,a person is jumping on a of of a of two men are ,0.80
,a woman is talking to the camera in a room ,0.72
,a young asian man is by a large and is by a person is singing ,0.90
,a man is down a water and a man and it it it it it ,0.95
,a woman sits in a chair at a chair and falls down the day ,0.83
,a young young man wearing a red and white pants are walking on a street ,0.89
,a man dressed in a red dress with a dress is dancing ,0.72
,a group of people are on a stage ,0.57
,outside in a grassy of a dog is to a fence ,0.72
,a young woman is standing on a stage with a stage ,0.73
,a man in a city street in a city ,0.65
,a young man is singing in a stage ,0.57
,a man is a a in a of a ,0.63
,a person in a kitchen a woman is ,0.69
,a person is running in a street at the ,0.62
,a man in a white shirt and a white jumps with a house ,0.87
,a a ball is running at daytime ,0.49
,a woman is holding a on a a and a young man is singing and ,0.99
,a man is in a room ,0.45
,a person in a city of a building and a building and a building and ,0.94
,a person is jumping into the daytime ,0.56
,a young man wearing a red shirt and red and are standing on the side ,0.89
,a man in a running at a ,0.51
,a person is down a ball with a ball ,0.66
,a black man with a black and is by a ,0.82
,a young man wearing a white shirt and walking the the of the the day ,1.00
,a young man wearing a black jacket and a black is dancing in a room ,0.87
,a man is by a young man with a room ,0.63
,a black and white shirt is holding a man ,0.66
,three young men dressed in a black and a white is singing ,0.74
,a person on a toy ,0.40
,a young men in a black and is dancing in the air in the air ,0.90
,a white man wearing white shirt and white shirt is talking at a table ,0.88
,a boy is talking to a young girl in a ,0.66
```

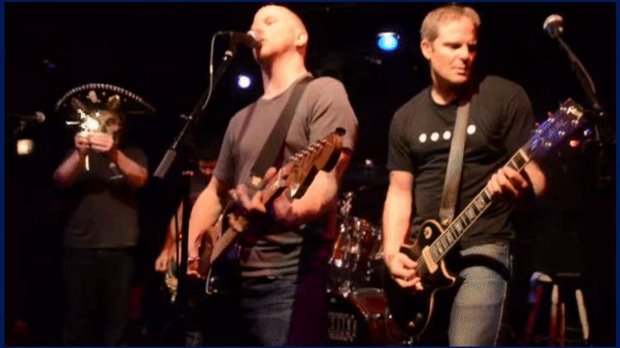
RESULTS:



a cat in a room

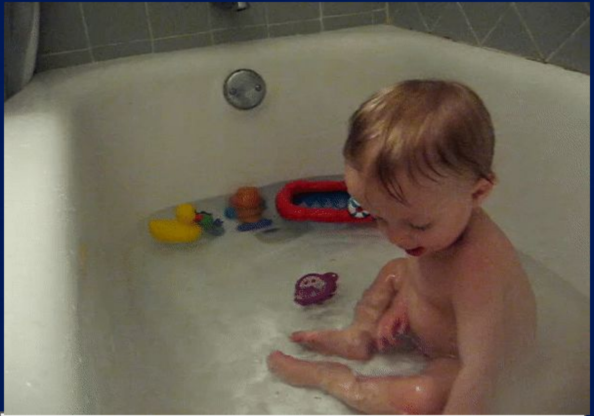


a group of men are standing in a room

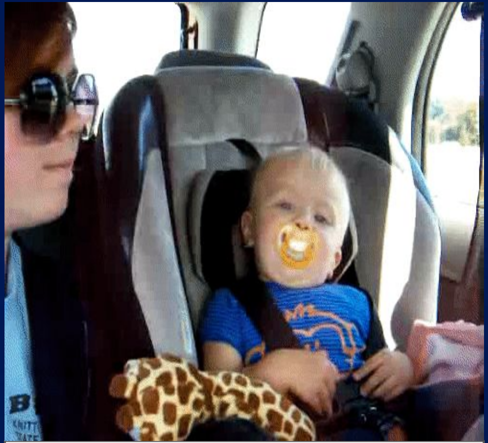


man is singing on a stage

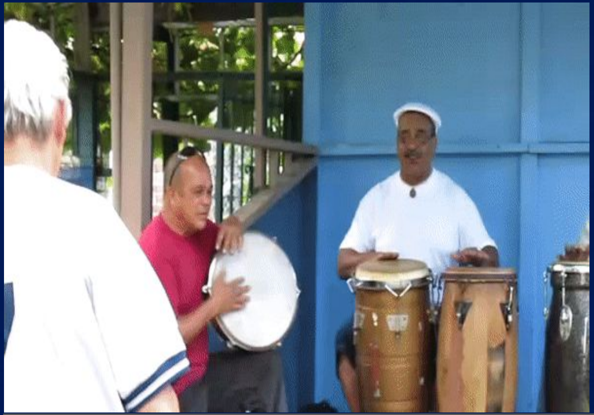
RESULTS:



a boy in the bathtub

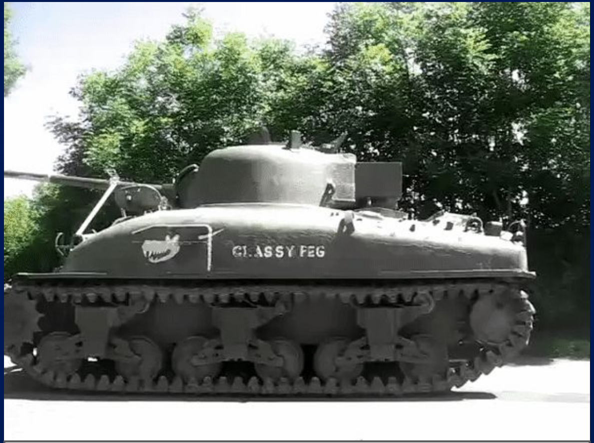


a boy is sitting on a car seat of a car

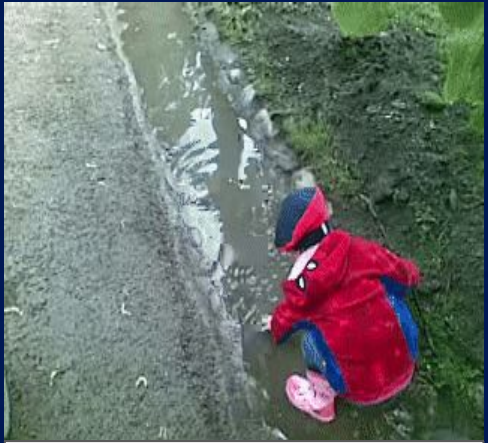


A man is talking to a crowd

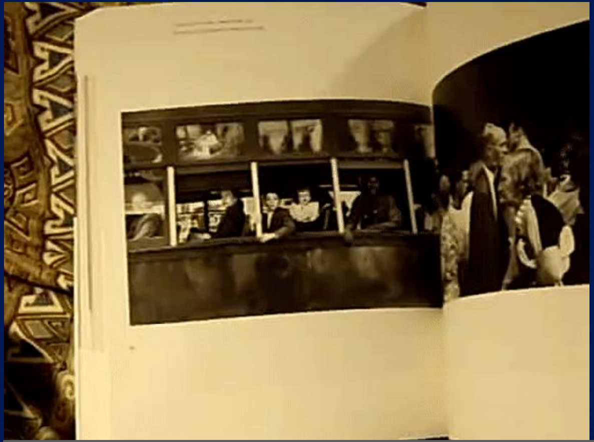
RESULTS :



a car on the road



a young man wearing a red shirt is jumping on a road



A video of a video of a person

REFERENCES:

> > Sequence to Sequence : Video to Text
https://arxiv.org/abs/1505.00487

>>
https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html

Other useful references:
>> https://arxiv.org/abs/1411.4555
>> https://arxiv.org/abs/1708.02043

Thank You