Video Captioning summary

C. Aslanidou, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 1.2
Date: September 2021

Artificial Intelligence &
Information Analysis Lab

Contents

- Video Captioning
  - Caption
    - Captioning Types, Methods, and Styles
  - Approaches
  - Methodology of approaching Video Captioning problem
  - Evaluation Metrics
  - Datasets
  - Future Directions
  - Video Captioning by Adversarial LSTM (an example)
  - Deep Learning for Video Captioning (an example)
  - References

Video Captioning

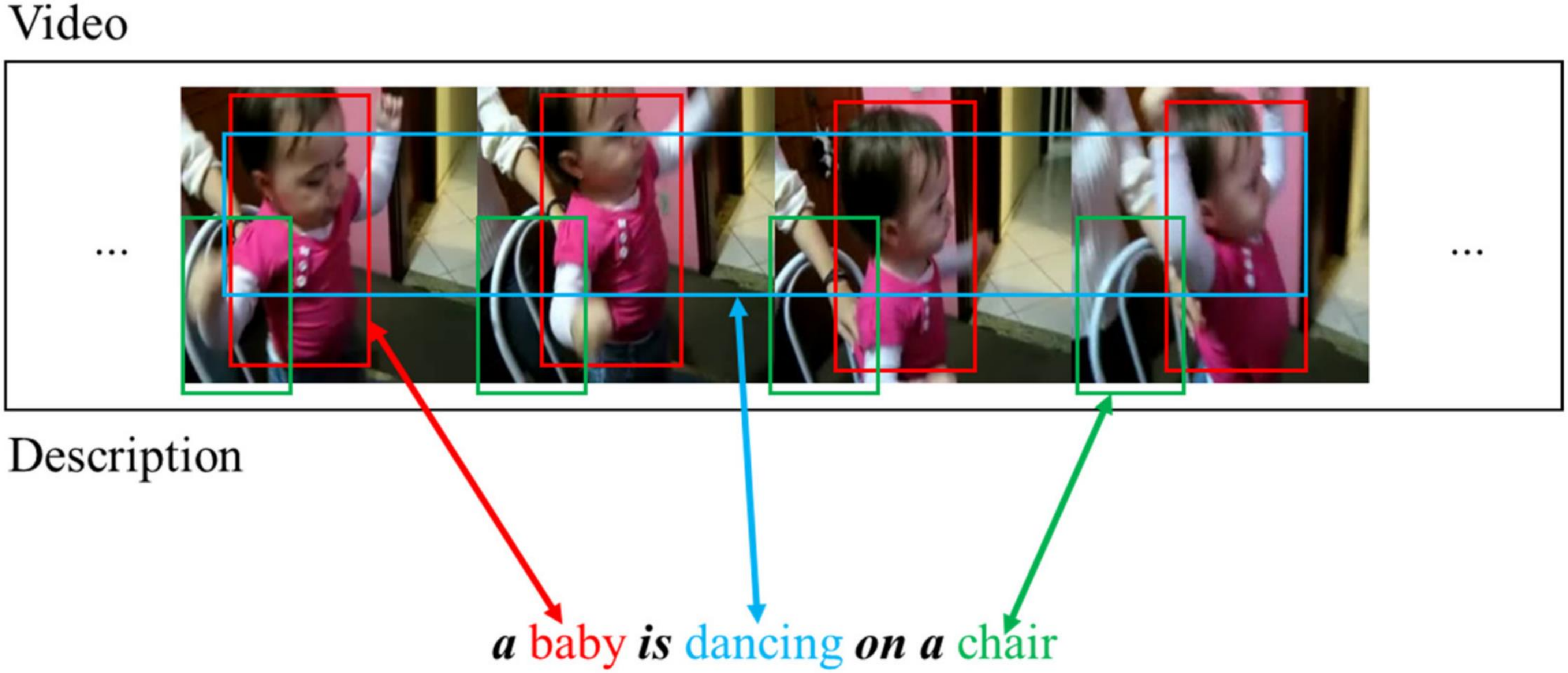


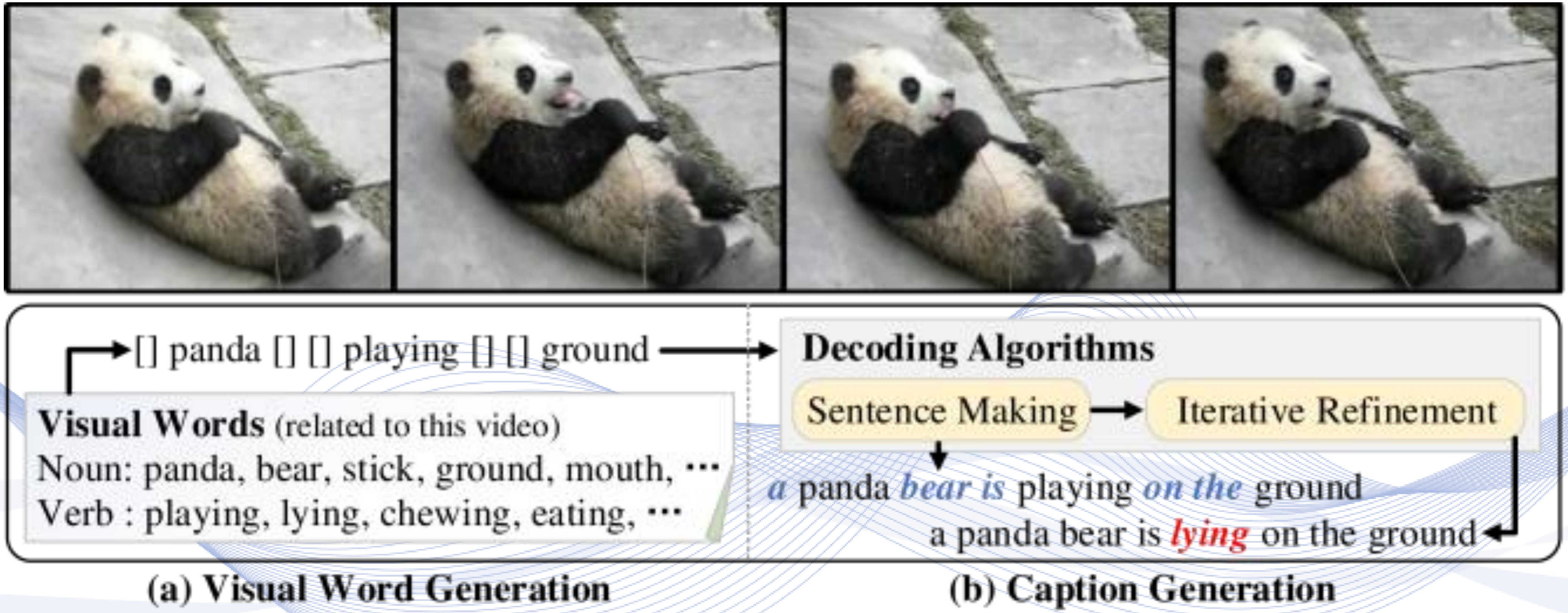
Image from mdpi

Video Captioning

Video captioning, has been showing increasingly strong potential in computer vision.

The primary challenges of this research lie in two aspects: **adequately** extracting the information from the video sequences and **generating grammar-correct sentences** easy for the human to understand. [YANG2018]

Video Captioning



[] panda [] [] playing [] [] ground
Visual Words (related to this video)
Noun: panda, bear, stick, ground, mouth, ···
Verb : playing, lying, chewing, eating, ···
Decoding Algorithms
Sentence Making
Iterative Refinement
a panda bear is playing on the ground
a panda bear is lying on the ground
(a) Visual Word Generation
(b) Caption Generation

Image from arxiv-vanity

Artificial Intelligence & Information Analysis Lab

Video Captioning

The early research for generating video descriptions mainly focused on extracting useful information e.g., object, attribute, and preposition, from given video content.

The aim is to generate more precise words to describe the objects in the video.

Video Captioning

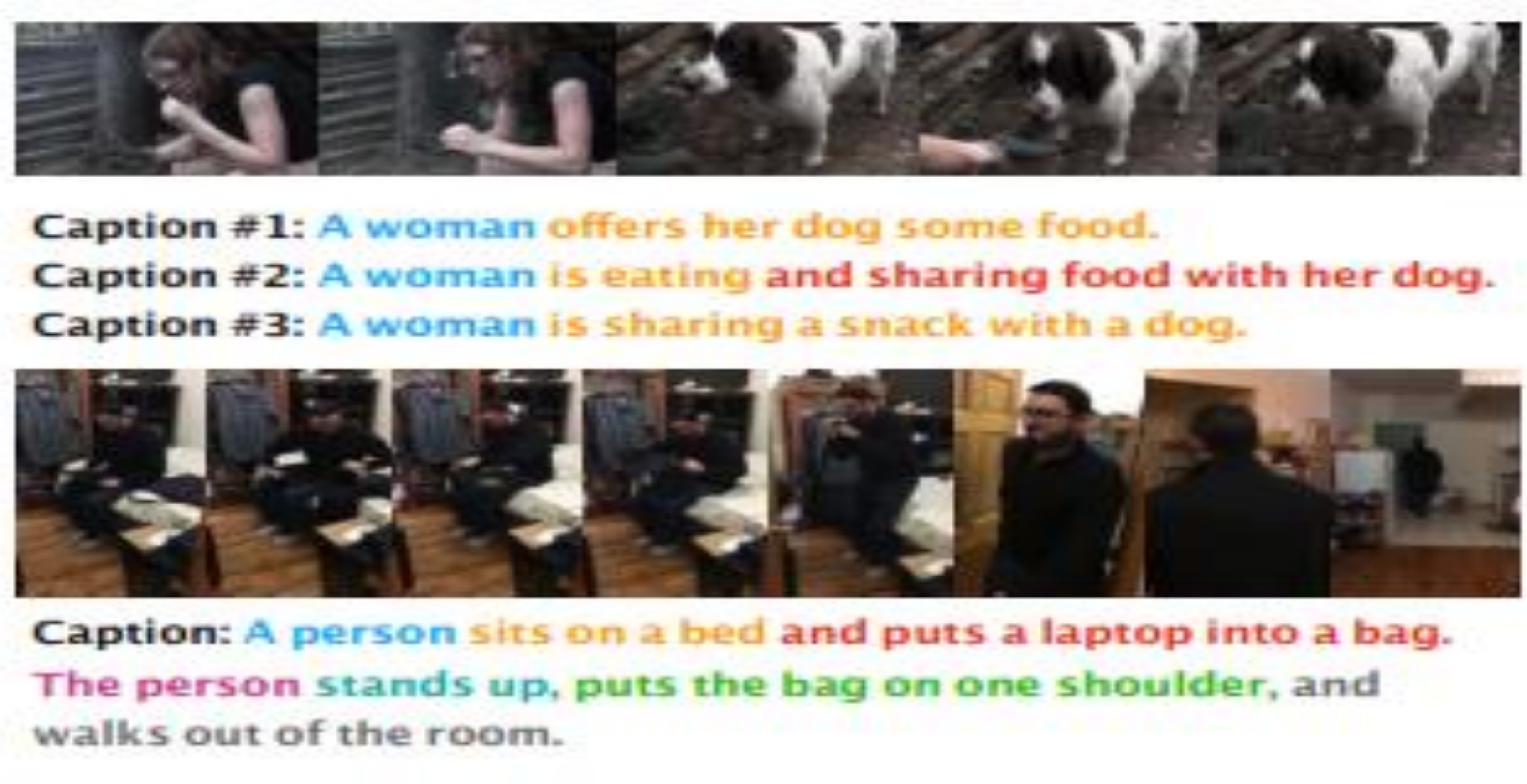
Deep learning methodologies have increased great focus towards video processing because of their better performance and the high-speed computing capability.

Video Captioning

Based on the approaches proposed for video captioning till now, they can be classified into **two categories** namely:

- The **template-based** language model and [ZAC2012]
- The **sequence** learning model. [YAN2016]

Video Captioning



Caption #1: A woman offers her dog some food.

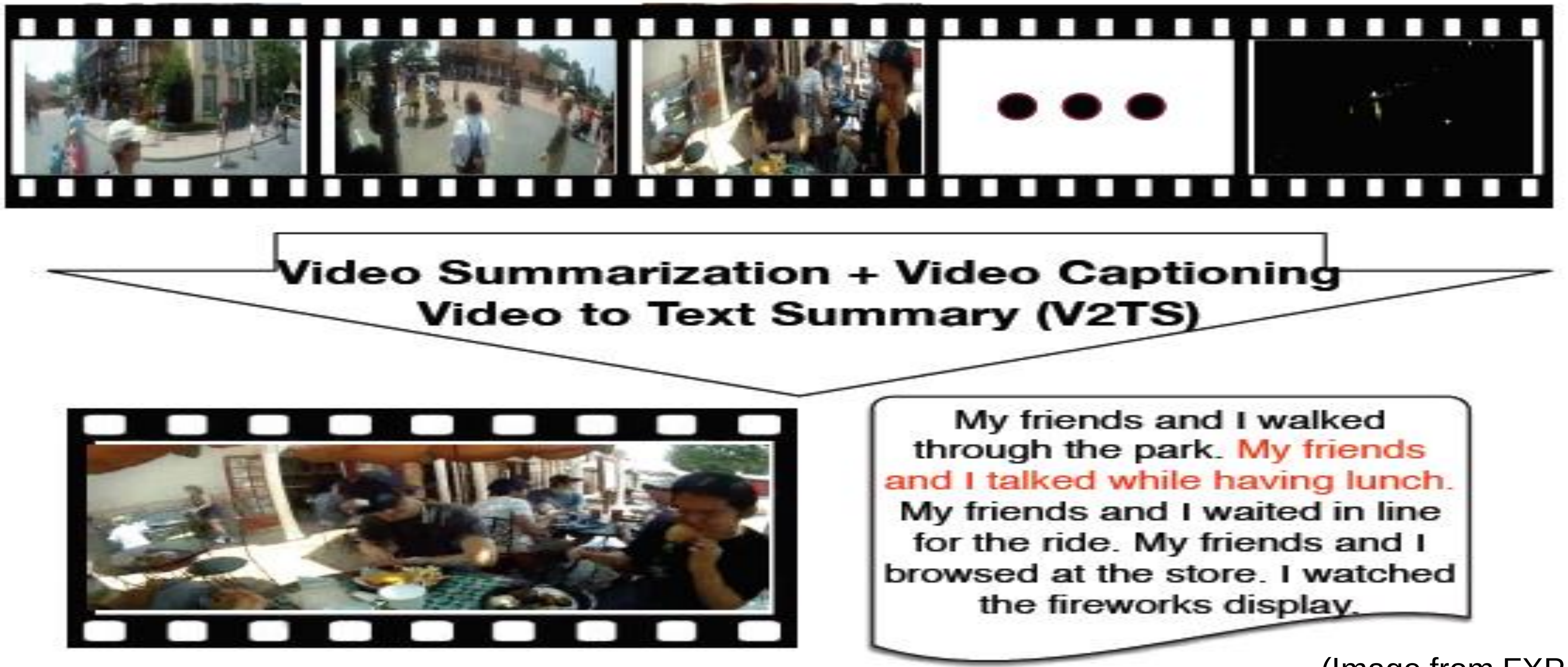
Caption #2: A woman is eating and sharing food with her dog.

Caption #3: A woman is sharing a snack with a dog.

Caption: A person sits on a bed and puts a laptop into a bag. The person stands up, puts the bag on one shoulder, and walks out of the room.

Video Captioning (Image from ResearchGate

Video Captioning



Video Summarization + Video Captioning
Video to Text Summary (V2TS)

My friends and I walked through the park. My friends and I talked while having lunch. My friends and I waited in line for the ride. My friends and I browsed at the store. I watched the fireworks display.

(Image from FXPAL)

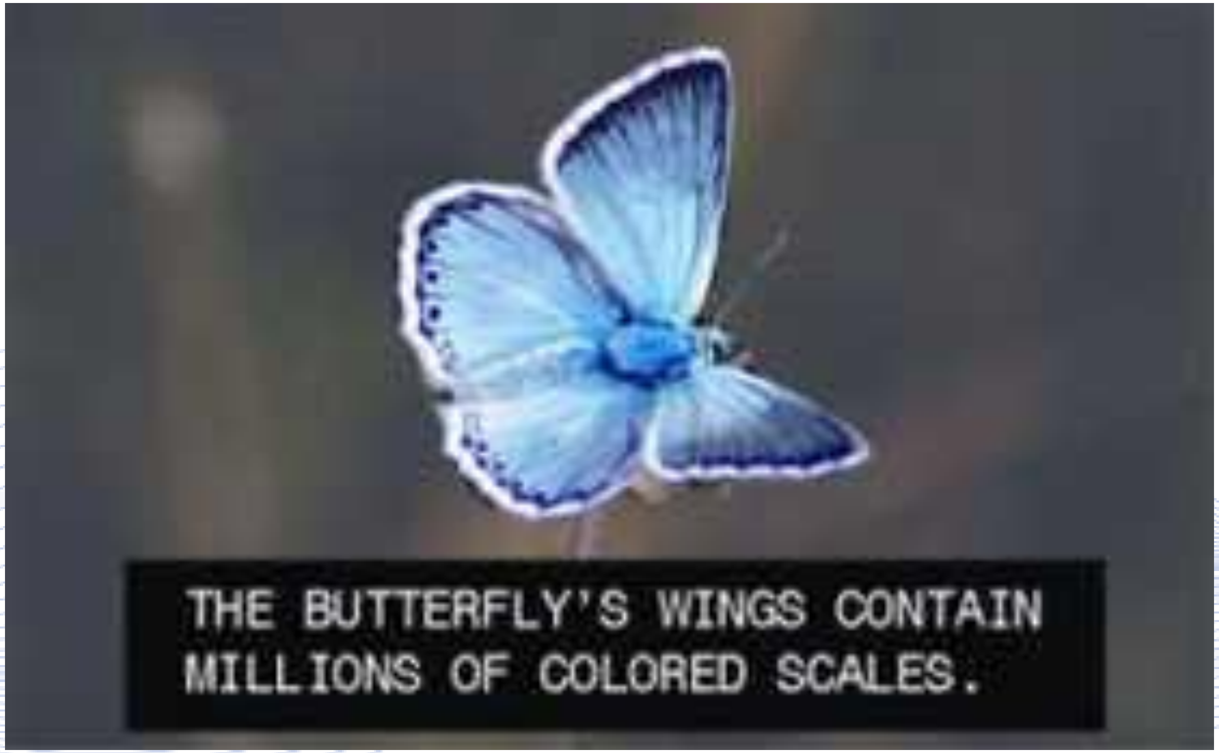
Video Captioning: Captioning Types, Methods, and Styles

- **Types**

Types vary according to how the captions appear, how they are accessed, and what information is provided. These include closed captions, subtitles, and subtitles for the deaf and hard of hearing. [HTT2021]

Video Captioning: Captioning Types, Methods, and Styles

Closed Captions

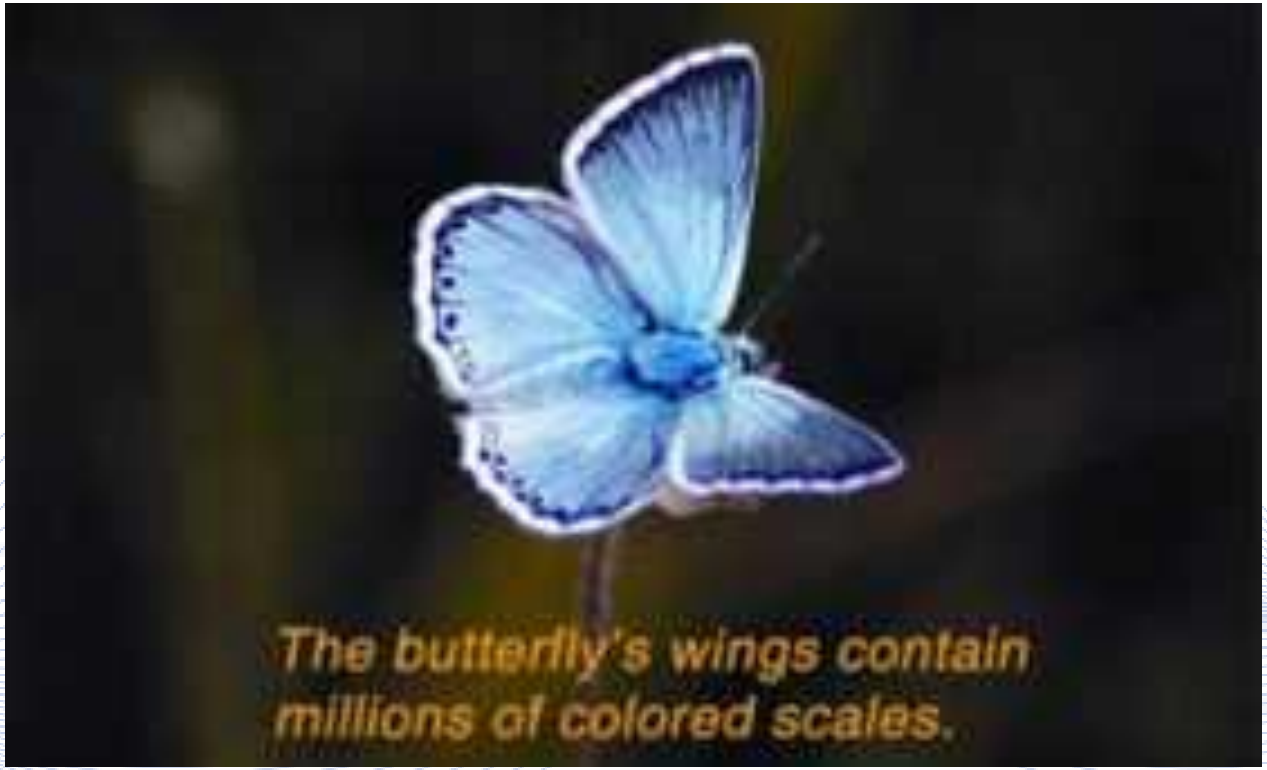


THE BUTTERFLY'S WINGS CONTAIN
MILLIONS OF COLORED SCALES.

Closed captions. (Image from https://dcmp.org/learn/38-captioning-types-methods-and-styles)

Video Captioning: Captioning Types, Methods, and Styles

Subtitles

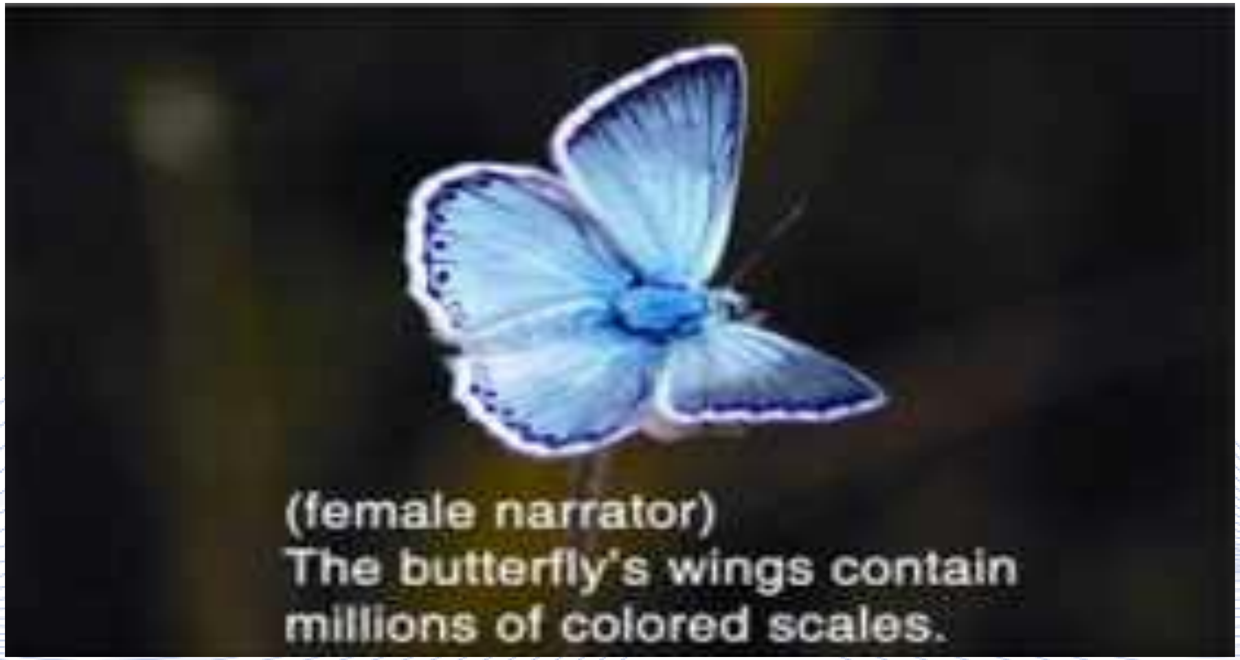


The butterfly's wings contain
millions of colored scales.

Subtitles. (Image from https://dcmp.org/learn/38-captioning-types-methods-and-styles)

Video Captioning: Captioning Types, Methods, and Styles

Subtitles for the Deaf and Hard of Hearing (SDH)



(female narrator)
The butterfly's wings contain
millions of colored scales.

Subtitles for the Deaf and Hard of Hearing. (Image from https://dcmp.org/learn/38-captioning-types-methods-and-styles)

Video Captioning: Captioning Types, Methods, and Styles

- **Methods**

  Methods vary according to when the captions are created and displayed.

  These include **off-line and on-line**. [HTT2021]

Video Captioning Approaches

The background of video **captioning approaches** can be divided into **three phases**:

- The **classical video captioning** approach phase involves the detection of entities of the video (such as object, actions and scenes) and then map them to a predefined templates. [LIU2019}

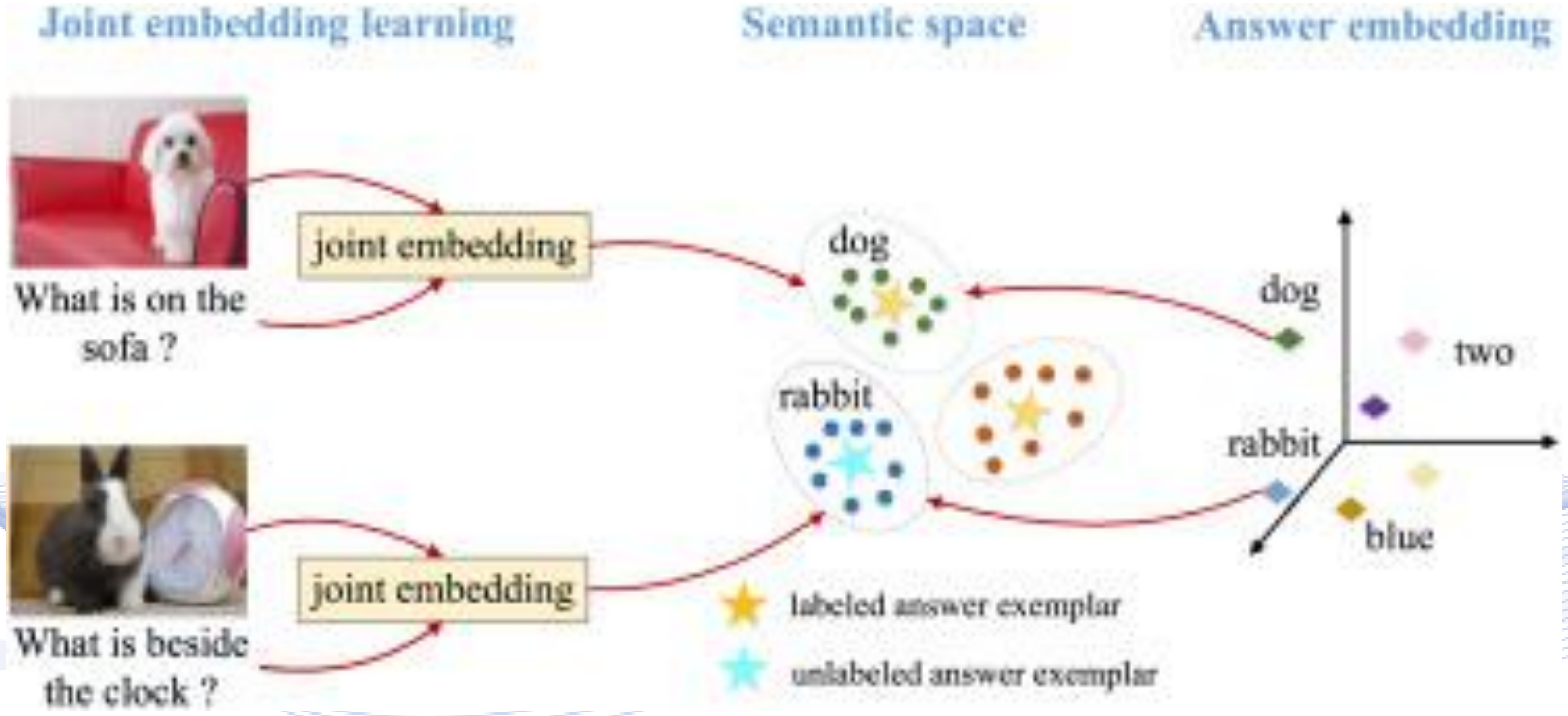
Video Captioning Approaches

- The **statistical methods** phase, in which the video captioning problem is addressed by employing statistical methods.

- The last one is **deep learning** phase. In this phase, many state-of-the-art video captioning frameworks have been proposed and it is believed that this phase has a capability of solving the problem of automatic open domain video captioning. [LIU2019}

Methodology of approaching Video Captioning problem

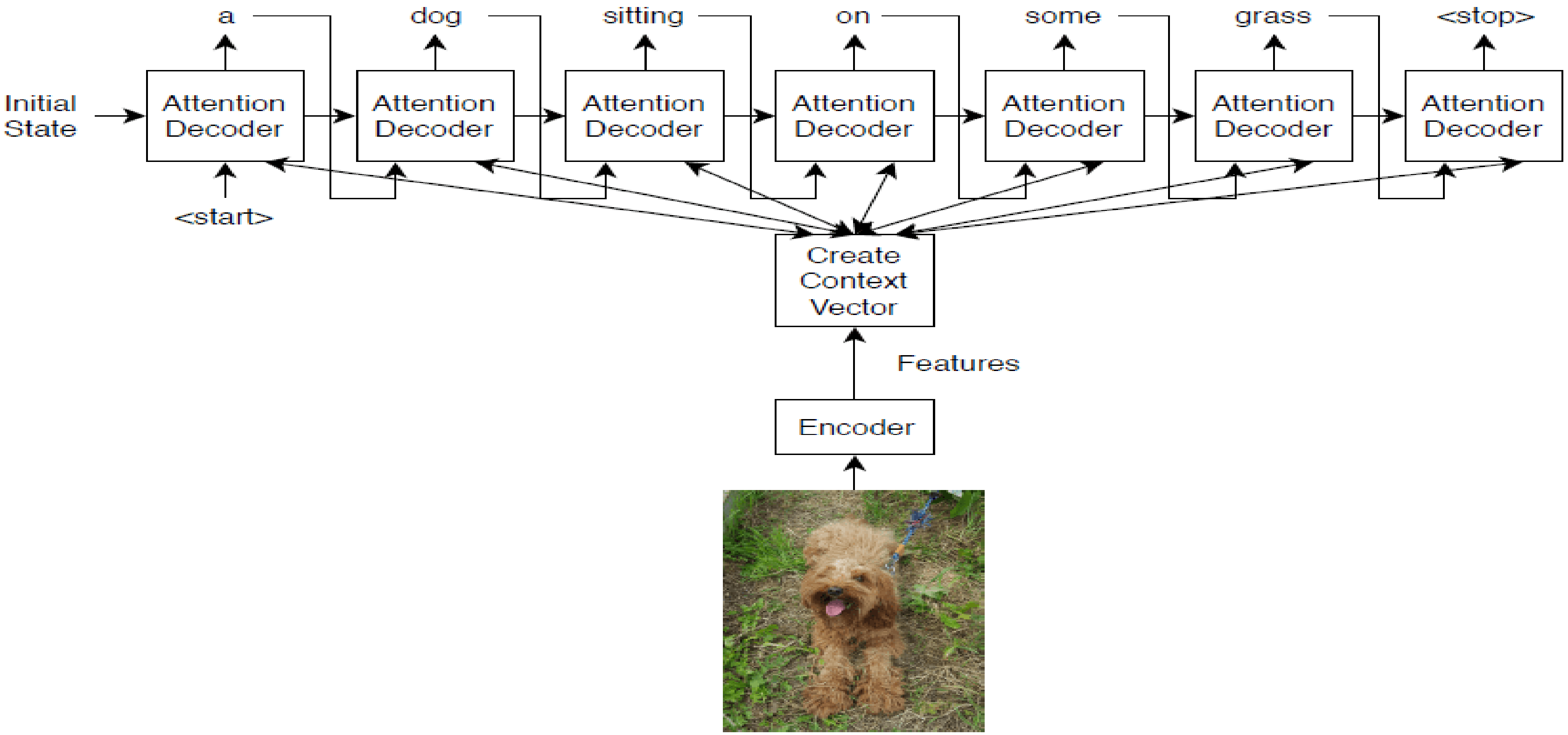
Joint Embedding



(Image from ScienceDirect)

Methodology of approaching Video Captioning problem

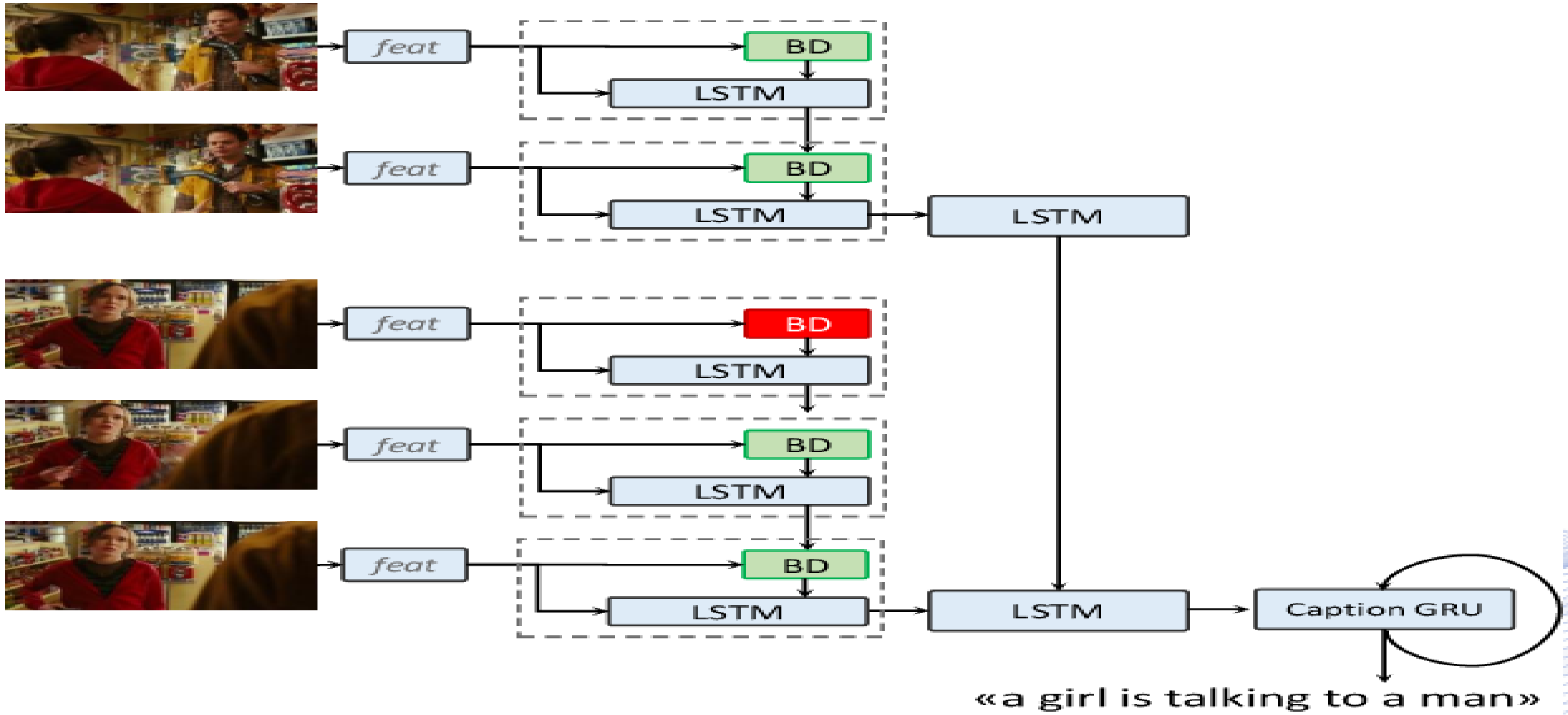
Encoder-Decoder Mechanisms: Attention Mechanism



(Image from MathWorks)

Methodology of approaching Video Captioning problem

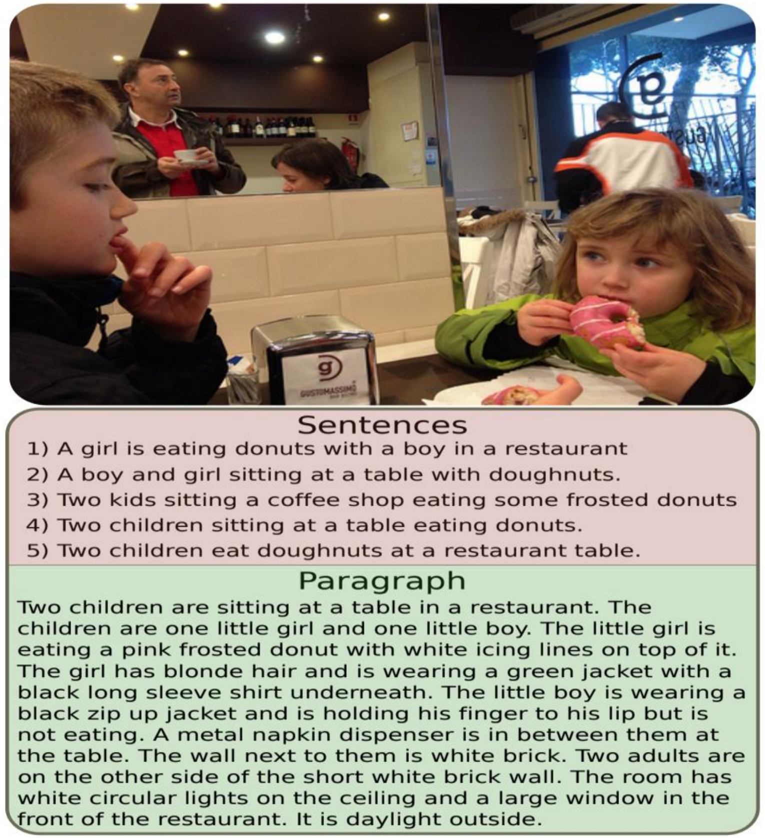
Encoder-Decoder Mechanisms: Hierarchical Neural Encoder



(Image from SemanticScolar)

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms: Paragraph Description



Sentences

1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

Paragraph

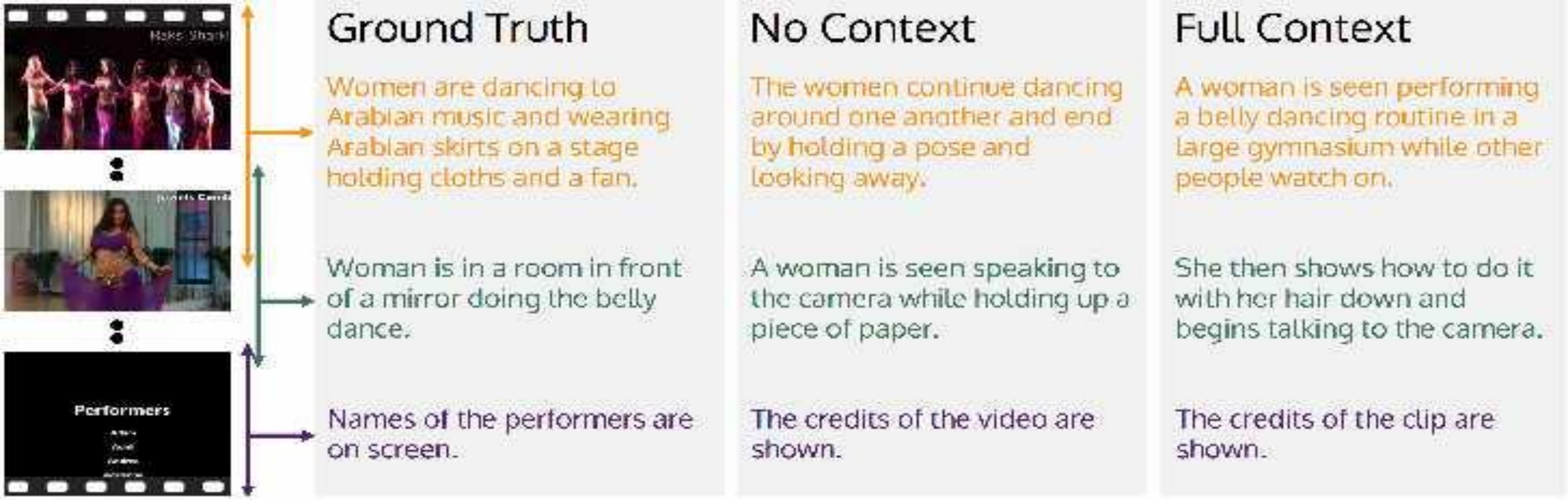
Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

(Image from Medium.com)

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**



Ground Truth	No Context	Full Context
Women are dancing to Arabian music and wearing Arabian skirts on a stage holding cloths and a fan.	The women continue dancing around one another and end by holding a pose and looking away.	A woman is seen performing a belly dancing routine in a large gymnasium while other people watch on.
Woman is in a room in front of a mirror doing the belly dance.	A woman is seen speaking to the camera while holding up a piece of paper.	She then shows how to do it with her hair down and begins talking to the camera.
Names of the performers are on screen.	The credits of the video are shown.	The credits of the clip are shown.

(Image from https://cs.stanford.edu/people/ranjaykrishna/densevid/)

Video Captioning: Evaluation Metrics

Video captioning result is evaluated based on correctness as natural language and relevance of semantics to its respective video.

The following are widely used evaluation metrics that concern the aspects.

Video Captioning: Evaluation Metrics - SVO

SVO Accuracy is used in early works to measure whether the generated SVO (Subject, Verb, Object) triplets cohere with ground truth.

The purpose of this evaluation metrics is to focus on matching of broad semantics and ignore visual and language details. [DON2014], [LIU2019]

Video Captioning: Evaluation Metrics -BLEU

BLEU is one of the most popular metrics in the field of machine translation. The idea is measuring a numerical translation closeness between two sentences by computing geometric mean of n-gram match counts. As a result, it is sensitive to position mismatching of words.

Also, it may favor shorter sentences, which makes it hard to adapt to complex contents. [PAR2017]

Video Captioning: Evaluation Metrics - ROUGE

ROUGE is similar to BLEU score in the sense that they measure the n-gram overlapped sequences between the reference sentences and the generated ones. The difference is that ROUGE considers the n-gram occurrences in the total sum of the number of reference sentences while BLEU considers the occurrences in the sum of candidates. Since ROUGE metric relies highly on recall, it favors long sentences. [PAR2017]

Video Captioning: Evaluation Metrics - CIDER

CIDER is a metric to evaluate a set of descriptive sentences for an image, which measures the consensus between candidate captioning and the reference sentences provided by human annotators. Therefor, it highly correlates with human judgments. It is different from others in the sense that it captures saliency and importance, accuracy, and grammatical correctness, and importance, accuracy, and grammatical correctness. [PAR2017]

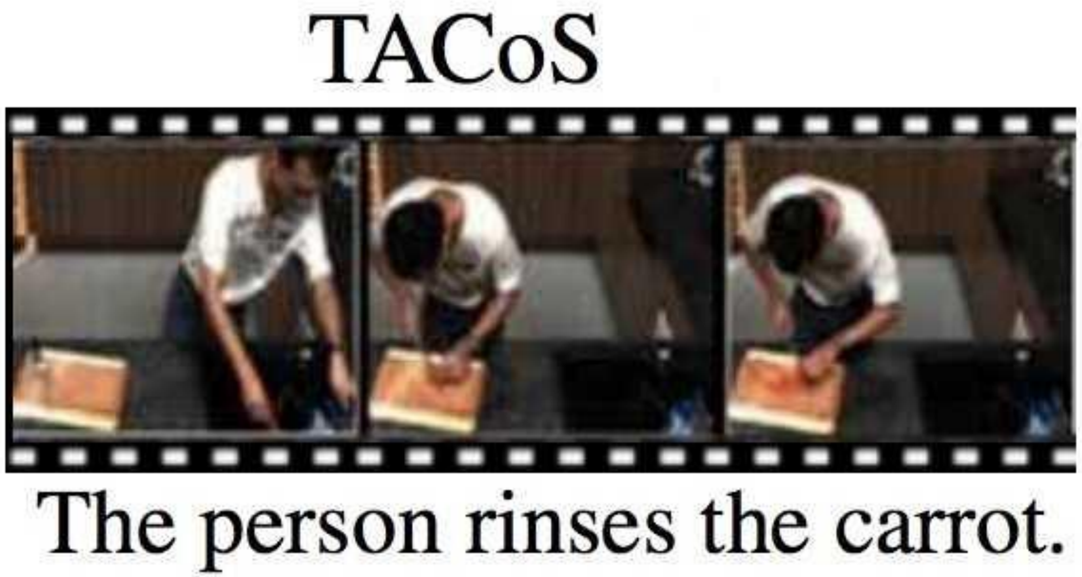
Video Captioning: Evaluation Metrics - METEOR

METEOR is computed based on the alignment between a given hypothesis sentence and a set of candidate reference. METEOR compares exact token matches, stemmed tokens, paraphrase matches, as well as semantically similar matches using WordNet synonyms. This semantic aspect of METEOR distinguishes it from others. METEOR is always better when the number of references is small. [PAR2017]

Video Captioning: Evaluation Metrics - F-Score

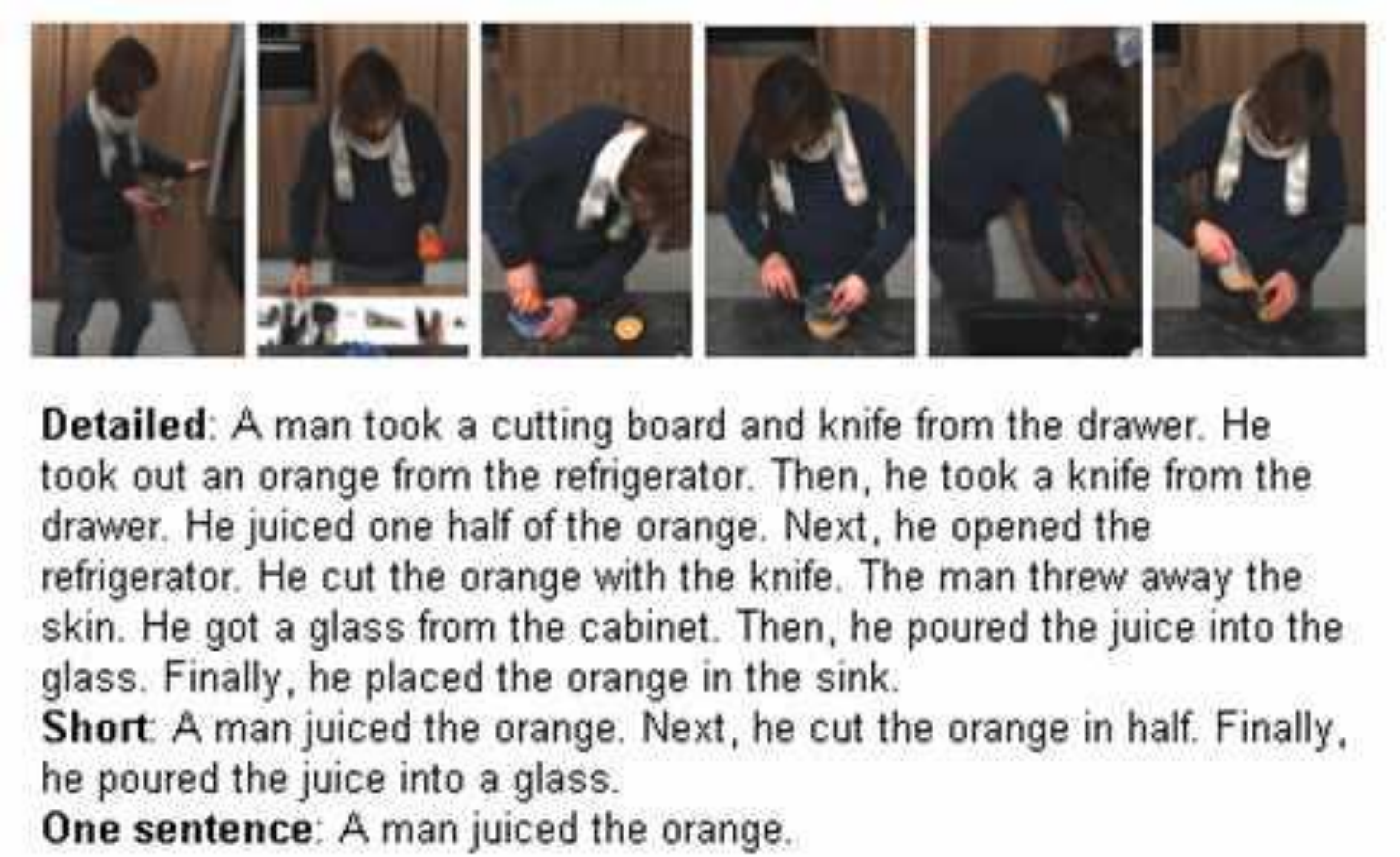
F-Score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’. It is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall. It is commonly used for evaluating information retrieval systems, and also for many kinds of machine learning models. [HTT2019]

Video Captioning: Datasets



Tacos dataset (Image from https://cove.thecvf.com/datasets/422)

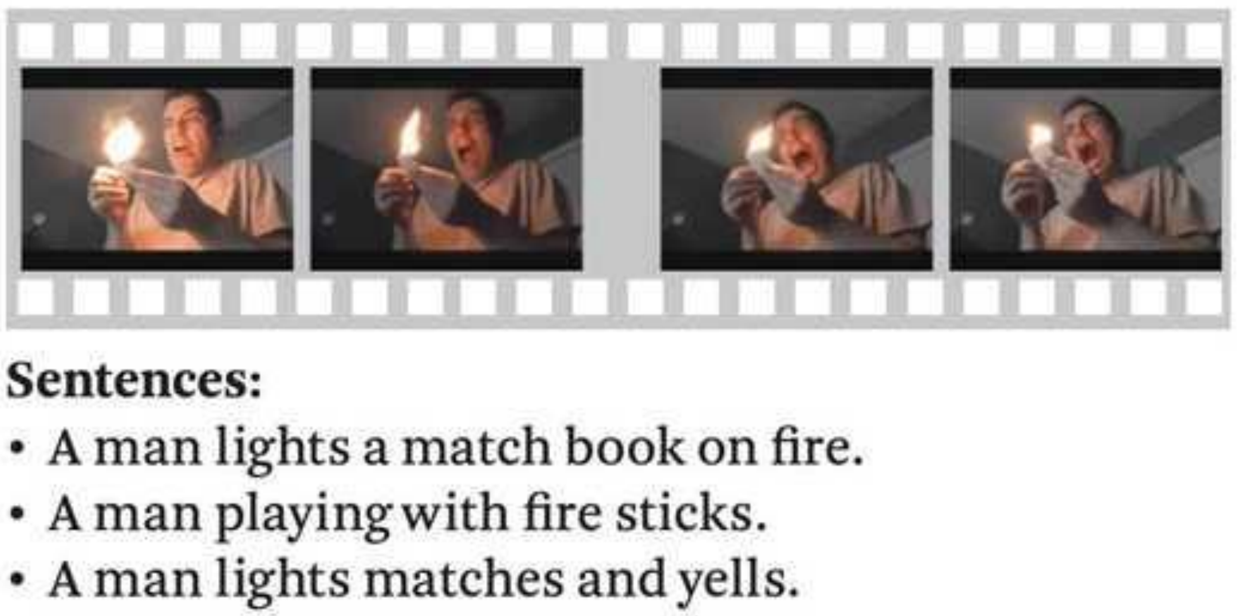
Video Captioning: Datasets



Detailed: A man took a cutting board and knife from the drawer. He took out an orange from the refrigerator. Then, he took a knife from the drawer. He juiced one half of the orange. Next, he opened the refrigerator. He cut the orange with the knife. The man threw away the skin. He got a glass from the cabinet. Then, he poured the juice into the glass. Finally, he placed the orange in the sink.
Short: A man juiced the orange. Next, he cut the orange in half. Finally, he poured the juice into a glass.
One sentence: A man juiced the orange.

https://www.mpi-inf.mpg.de/departments/computervision-and-machine-learning/research/vision-and-language/tacos-multi-level-corpus)

Video Captioning: Datasets

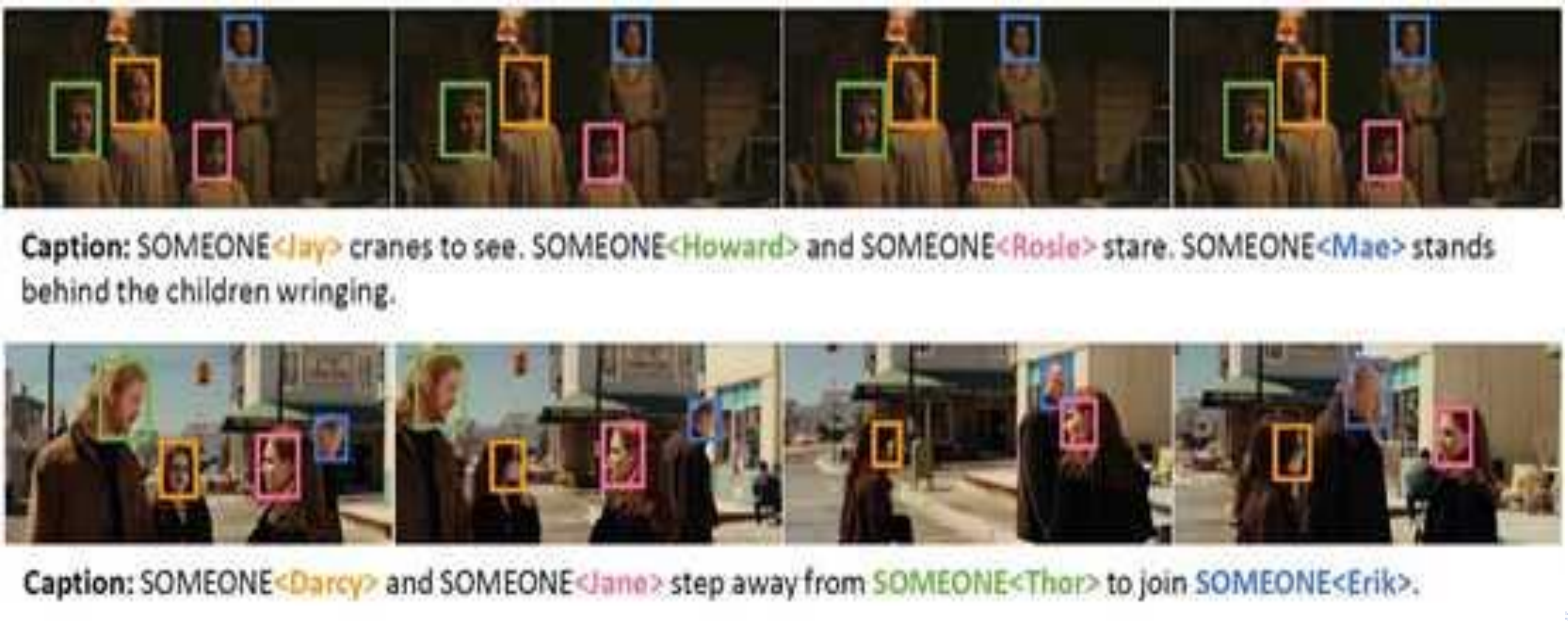


Sentences:

- A man lights a match book on fire.
- A man playing with fire sticks.
- A man lights matches and yells.

Microsoft Video Description Corpus (MSVD)
(Image from https://paperswithcode.com/dataset/msvd)

Video Captioning: Datasets

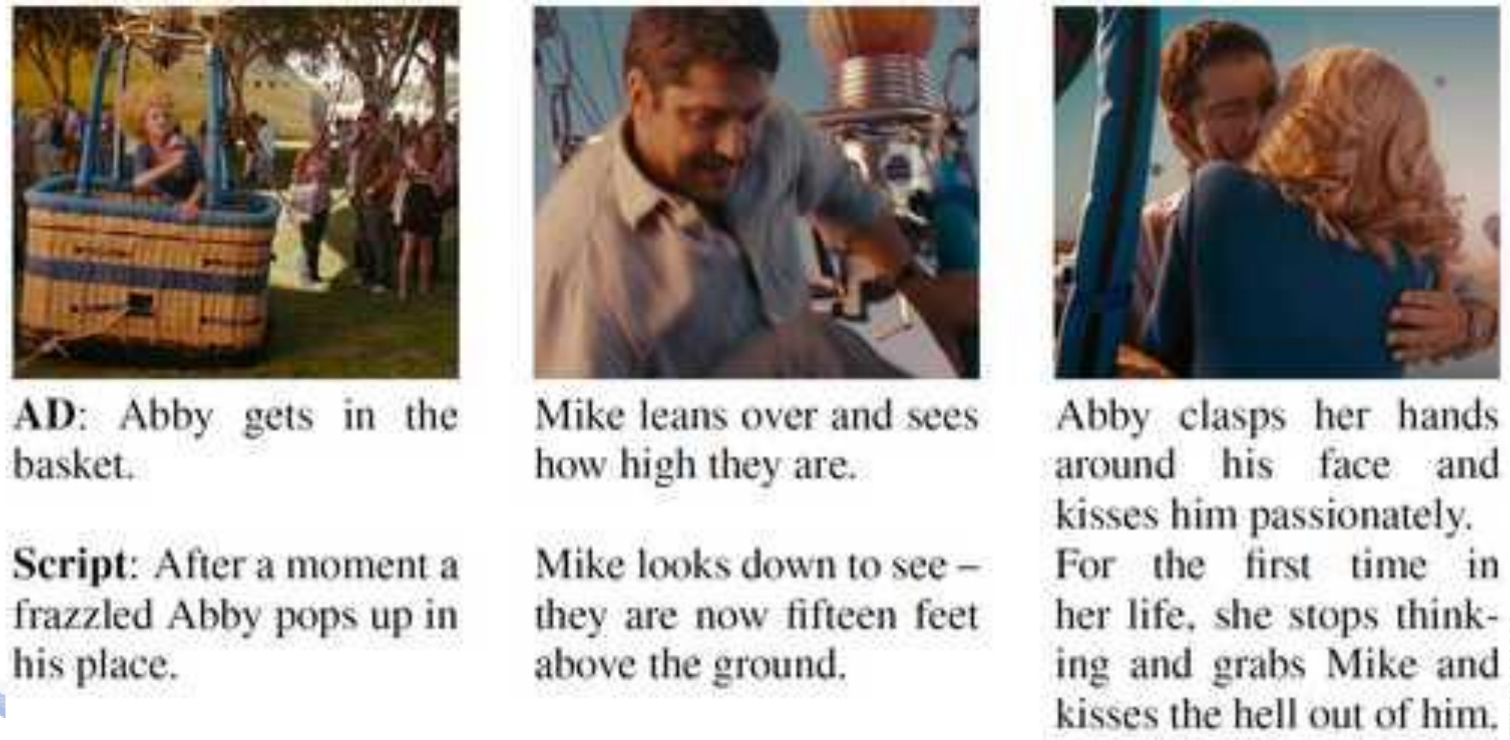


Caption: SOMEONE<Jay> cranes to see. SOMEONE<Howard> and SOMEONE<Rosie> stare. SOMEONE<Mae> stands behind the children wringing.

Caption: SOMEONE<Darcy> and SOMEONE<Jane> step away from SOMEONE<Thor> to join SOMEONE<Erik>.

Montreal Video Annotation Dataset (M-VAD) dataset
(Images from https://github.com/aimagelab/mvad-names-dataset)

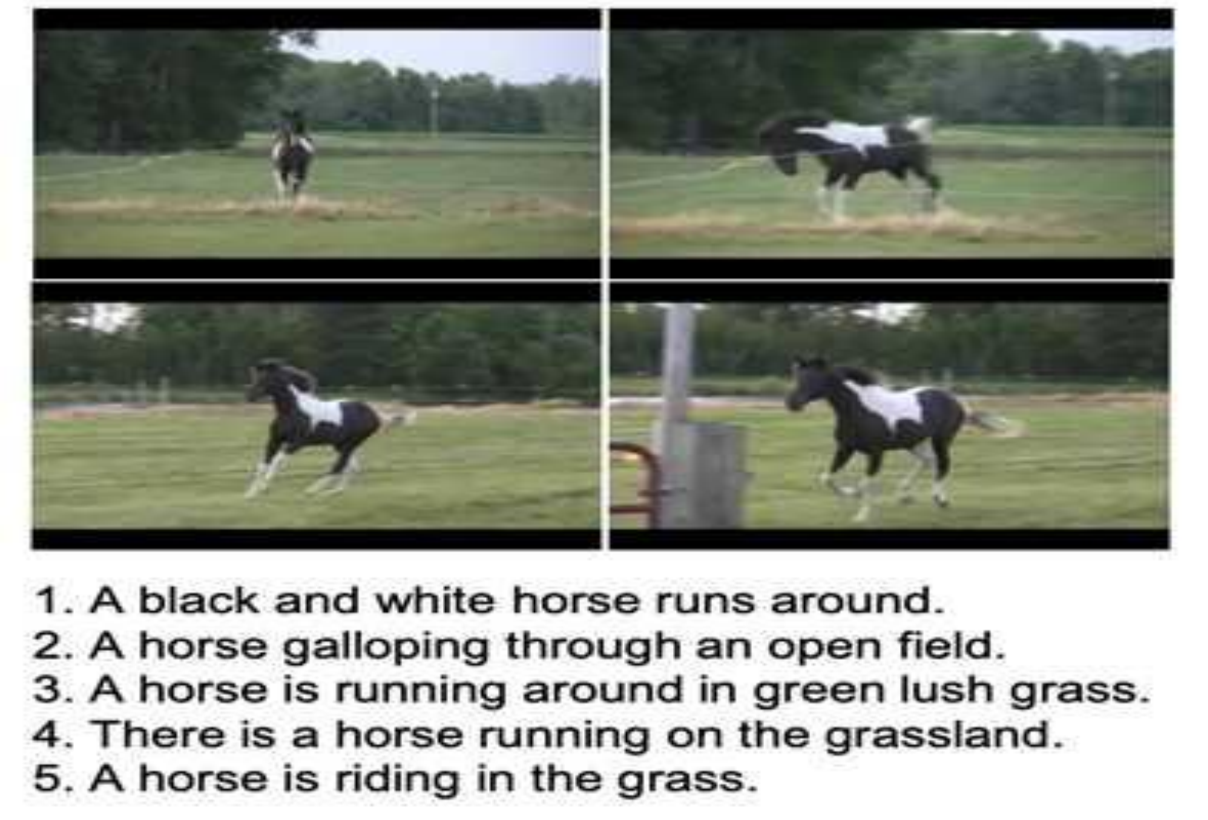
Video Captioning: Datasets



AD: Abby gets in the basket.	Mike leans over and sees how high they are.	Abby clasps her hands around his face and kisses him passionately.
Script: After a moment a frazzled Abby pops up in his place.	Mike looks down to see – they are now fifteen feet above the ground.	For the first time in her life, she stops thinking and grabs Mike and kisses the hell out of him.

MPII Movie Description Corpus (MPII-MD) dataset (Image from https://www.mpi inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/mpii-movie-description-dataset)

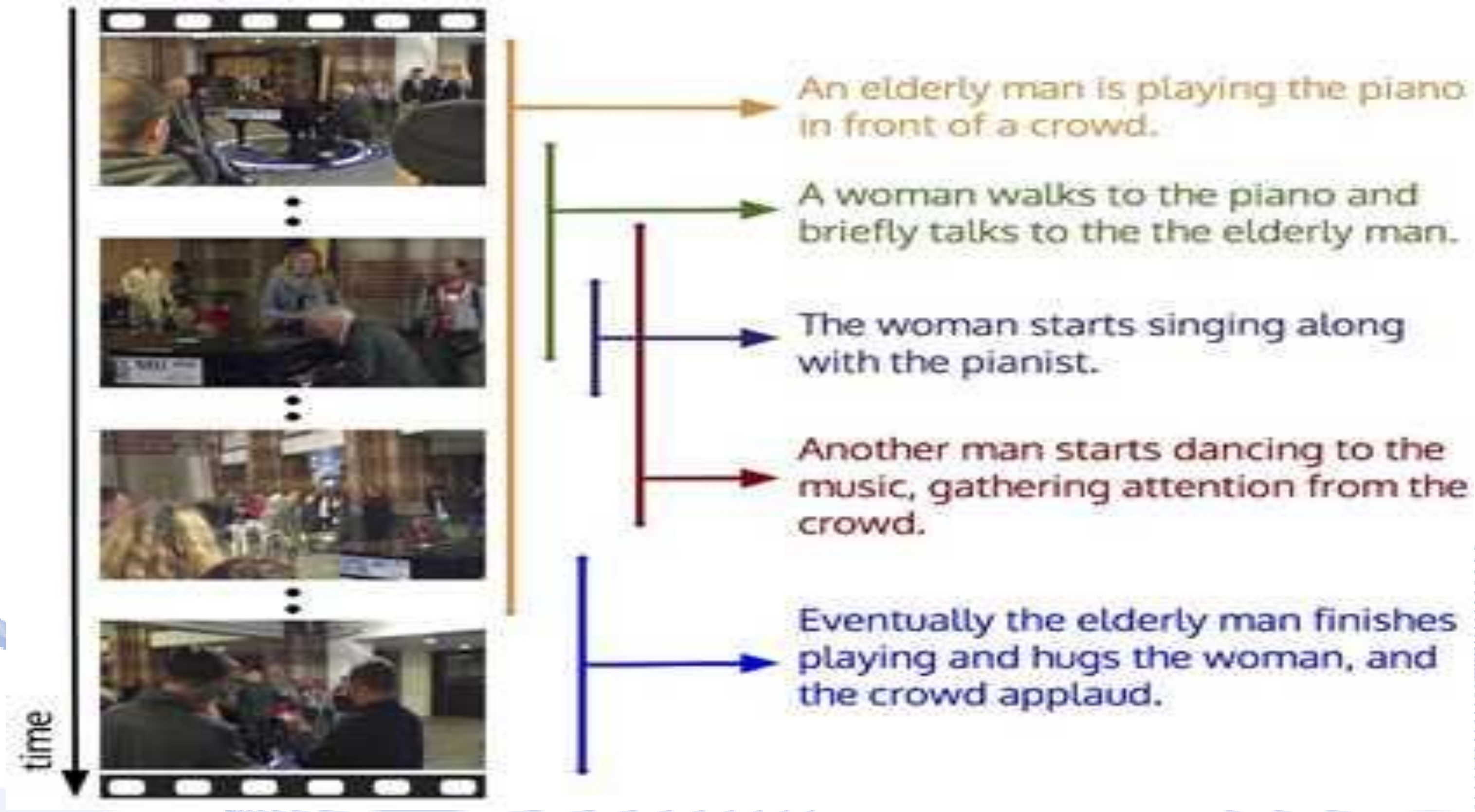
Video Captioning: Datasets



1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

MSR Video-to-Text (MSR-VTT) dataset. (Image from https://paperswithcode.com/dataset/msr-vtt)

Video Captioning: Datasets



An elderly man is playing the piano in front of a crowd.

A woman walks to the piano and briefly talks to the the elderly man.

The woman starts singing along with the pianist.

Another man starts dancing to the music, gathering attention from the crowd.

Eventually the elderly man finishes playing and hugs the woman, and the crowd applaud.

ActivityNet Captions dataset (Image from https://cs.stanford.edu/people/ranjaykrishna/densevid/)

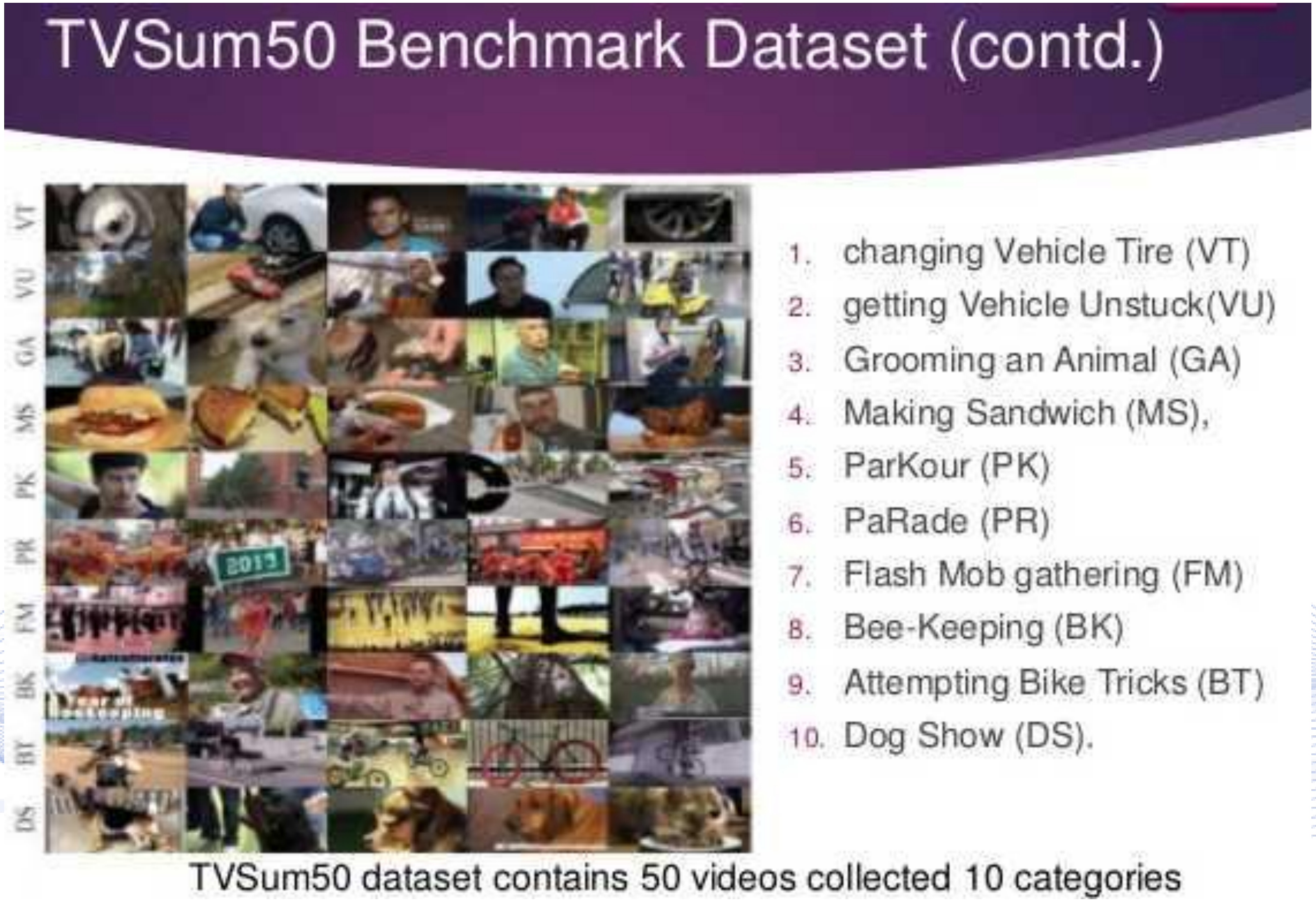
Video Captioning: Datasets

a) Air_Force_One

b) Play Ball

Sample of videos in the SumMe dataset (Image by ResearchGate)

Video Captioning: Datasets


VT
VU
GA
MS
PK
PR
FM
BK
BT
DS

TVSum50 Benchmark Dataset (contd.)

1. changing Vehicle Tire (VT)
2. getting Vehicle Unstuck(VU)
3. Grooming an Animal (GA)
4. Making Sandwich (MS),
5. ParKour (PK)
6. PaRade (PR)
7. Flash Mob gathering (FM)
8. Bee-Keeping (BK)
9. Attempting Bike Tricks (BT)
10. Dog Show (DS).

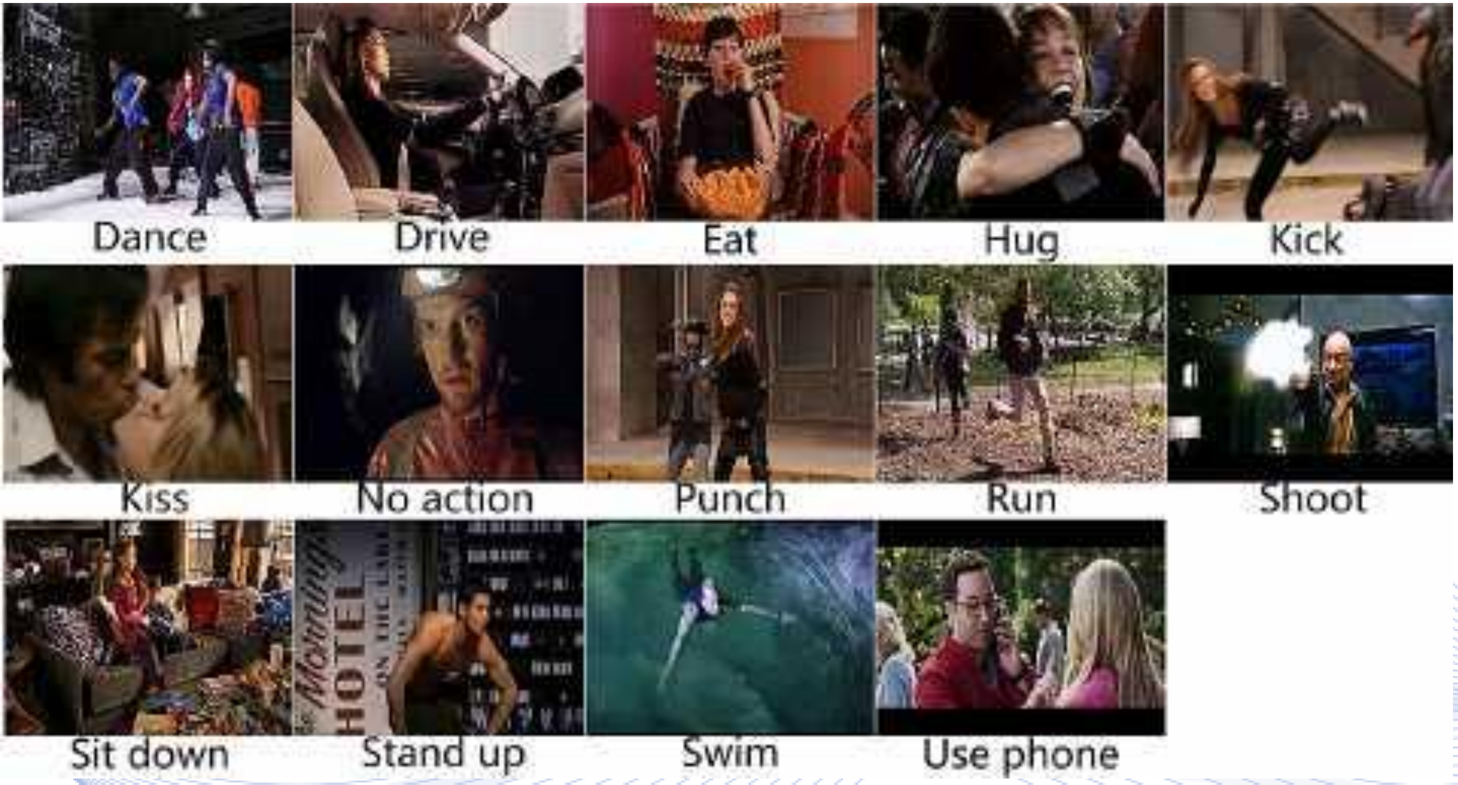
TVSum50 dataset contains 50 videos collected 10 categories

Sample of videos in the TVSum dataset (Image by ResearchGate)

Video Captioning: Datasets

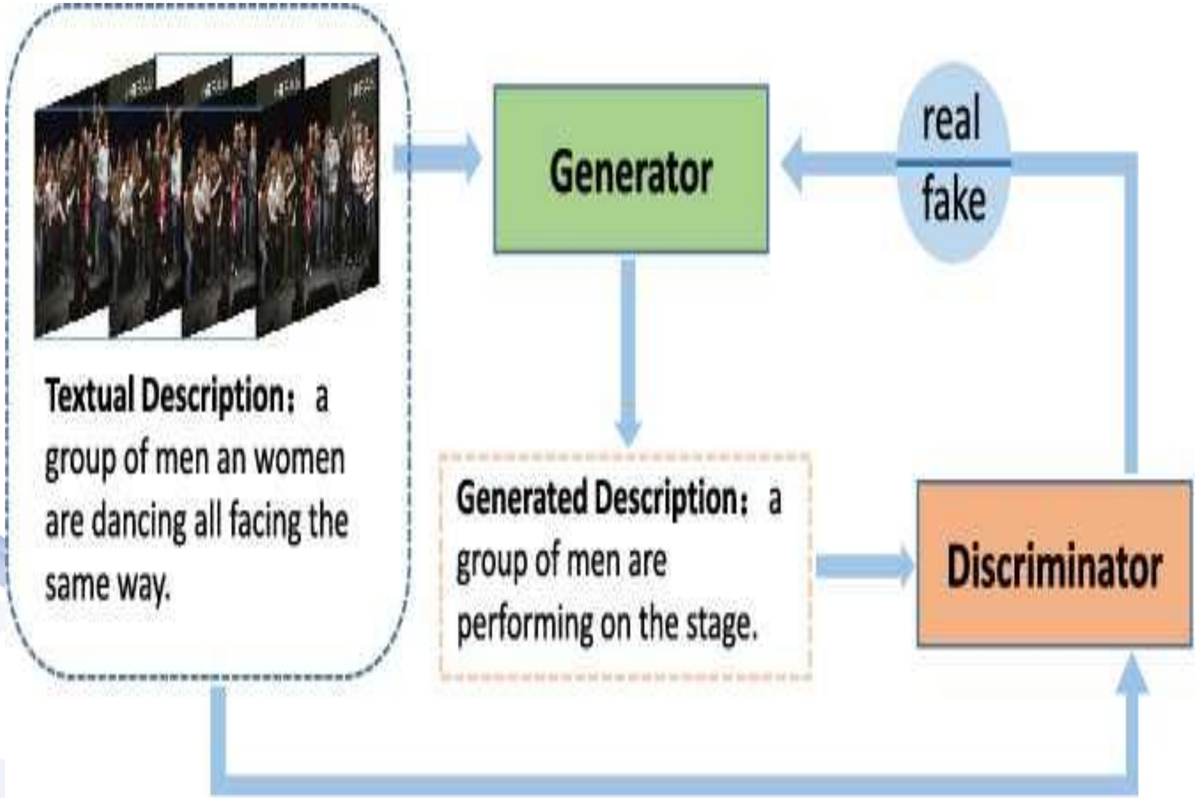
Sample of videos in the Hollywood2 dataset (Image by Researchgate)

Video Captioning: Datasets



Sample of videos in the Hollywood3D dataset (Image by ResearchGate)

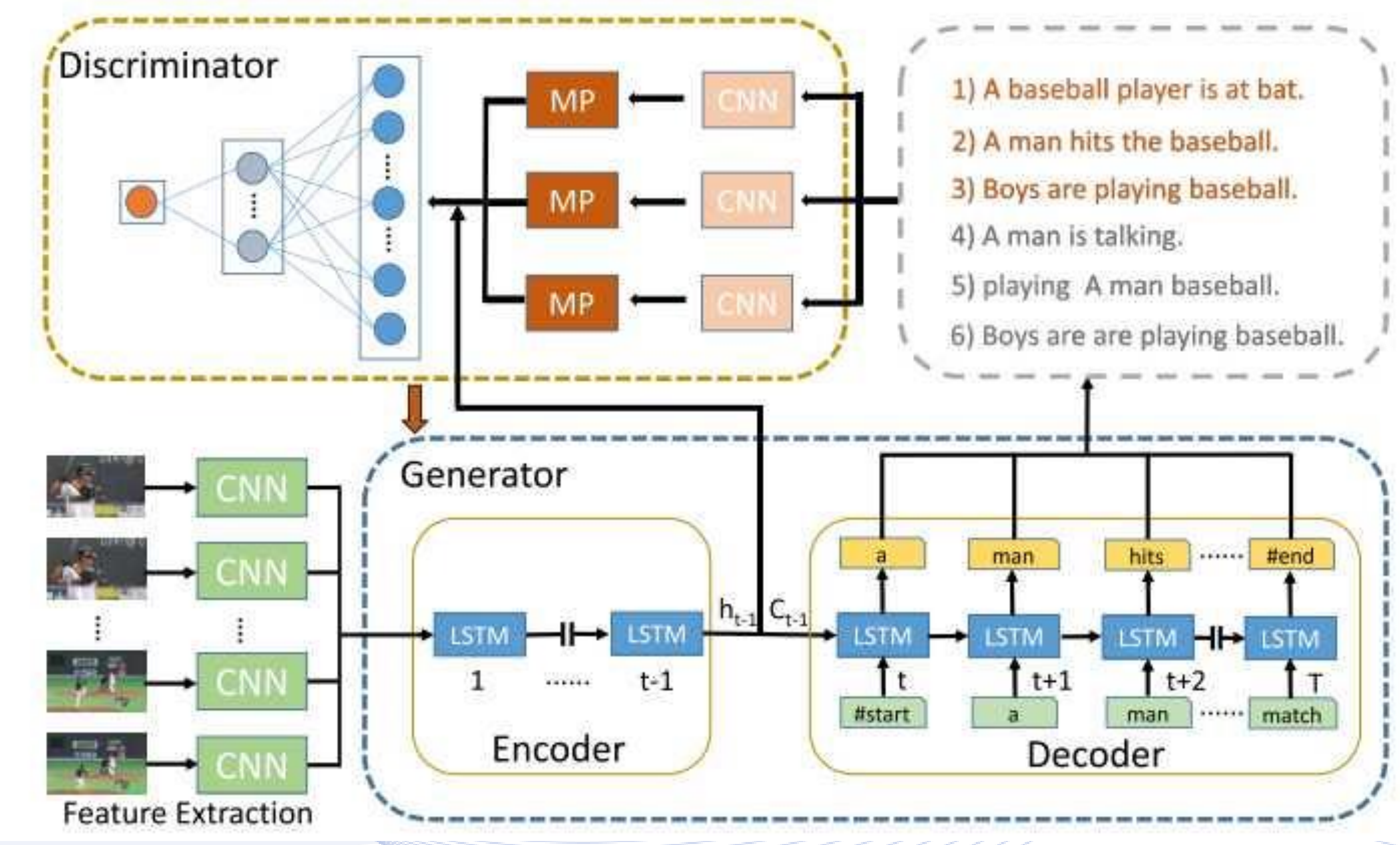
Video Captioning by Adversarial LSTM


Textual Description： a group of men an women are dancing all facing the same way.
Generator
real
fake
Generated Description： a group of men are performing on the stage.
Discriminator

An illustration of the modular structure of the proposed video captioning by an interplay of the generator G that generates text sentences and the discriminator D (adversary) that verifies the sentences. The optimization goal is that G deceives D, by generating sentences that are not distinguishable from reference sentences.

Dense-captioning events (Image from https://cs.stanford.edu/people/Video Captioning by Adversarial LSTM (Image from Semantic Scholar).

Video Captioning by Adversarial LSTM example



(Image from SemanticScholar).

LSTM-GAN incorporating a joint LSTMs with adversarial learning. The model consists of generative model and discriminative model. The generative model tries to generate a sentence for the video as accurately as possible, but the discriminative model tries to distinguish whether the input sentences is from reference sentence or generated sentences. The orange input sentences for discriminative model represent the reference sentences, otherwise badly constructed sentences or uncorrelated sentences generated by generative model. MP in the figure denotes the max-pooling. [ZHO2018]

Deep Learning for Video Captioning

Deep learning has achieved great successes in solving specific artificial intelligence problems recently. Substantial progresses are made on Computer Vision (CV) and Natural Language Processing (NLP). As a connection between the two worlds of vision and language, video captioning is the task of producing a natural-language utterance (usually a sentence) that describes the visual content of a video. The task is naturally decomposed into two sub-tasks.

Bibliography

[PIT2016] I. Pitas (editor), “Graph analysis in social media”, CRC Press, 2016.

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.

[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).

[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.

[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.

[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr