Video captioning: can neural networks describe a video?

Silvio Olivastri - Senior Data Scientist

{codemotion}

Online Tech Conference

- Italian edition -

23-24-25 Marzo, 2021

Index

1) Artificial Intelligence Challenges
2) Describing videos as combination of CV and NLP
3) Video Captioning
4) Benchmark Datasets
5) How to evaluate a Video Captioning model?
6) Metrics
7) Baseline models
8) State of the art models
9) Conclusions
10) References

Artificial Intelligence Challenges

- Reasoning
- Perception
- Motion and manipulation
- Planning
- Natural Language Processing
- Social Intelligence
- Learning
- General Intelligence
- Knowledge Representation

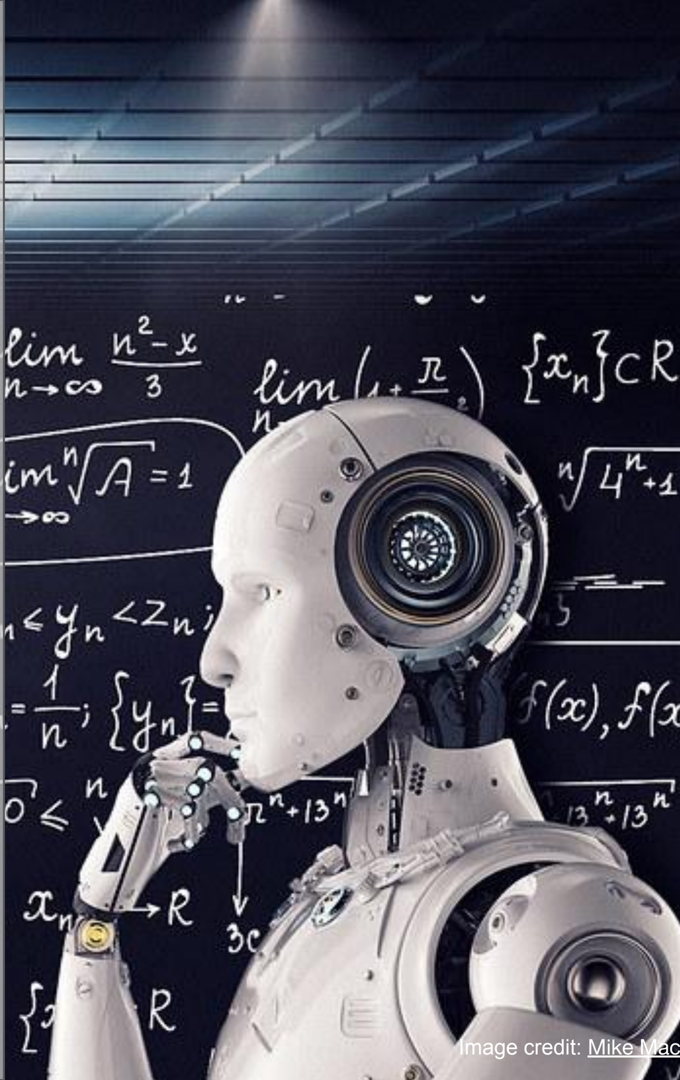
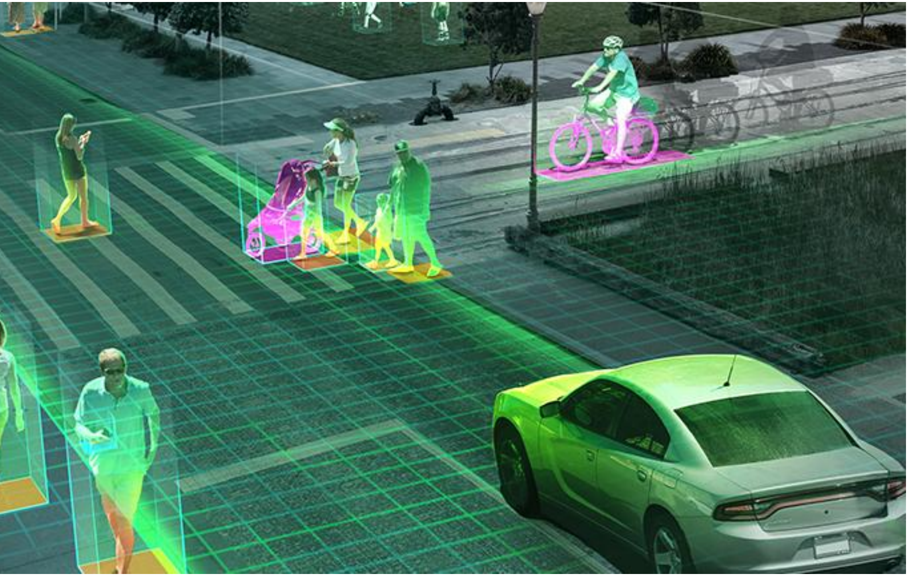


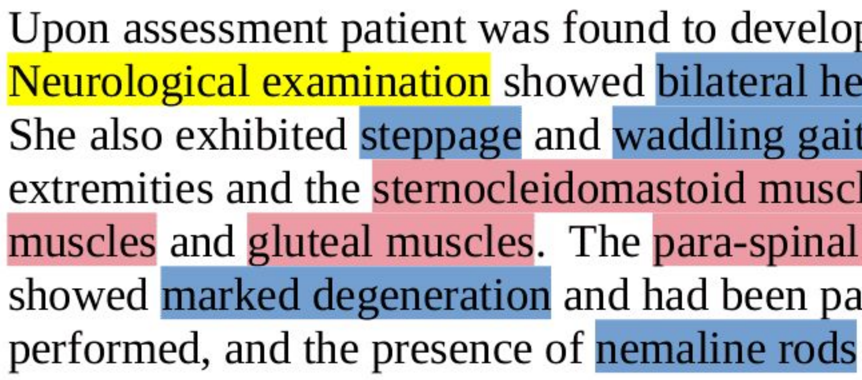
Image credit: Mike Mac

Artificial Intelligence Challenges

Machine perception is the ability to use input from sensors, like cameras or microphones, to deduce aspects of world. There are many famous applications such as object recognition or speech recognition.

Natural Language Processing (NLP) allows machines to read, understand and generate human language. Some applications are: sentiment analysis, information retrieval, machine translation and question answering.





Upon assessment patient was found to develo
Neurological examination showed bilateral he
She also exhibited steppage and waddling gait
extremities and the sternocleidomastoid muscl
muscles and gluteal muscles. The para-spinal
showed marked degeneration and had been pa
performed, and the presence of nemaline rods

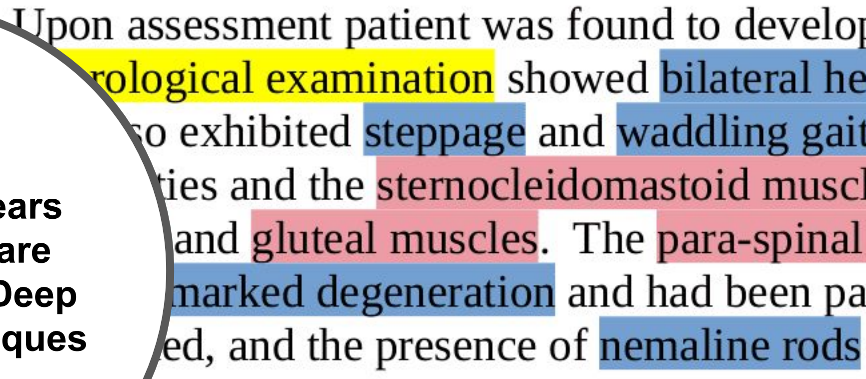
Image credit: nvidia.com, fastdatascience.com

Artificial Intelligence Challenges

Machine perception is the ability to use input from sensors, like cameras or microphones, to deduce aspects of world. There are many famous applications such as object recognition or speech recognition.

Natural Language Processing (NLP) allows machines to read, understand and generate human language. Some applications are: sentiment analysis, information retrieval, machine translation and question answering.

In the latest years both of them are dominated by Deep Learning techniques



Upon assessment patient was found to develop
rological examination showed bilateral he
so exhibited steppage and waddling gait
ties and the sternocleidomastoid muscl
and gluteal muscles. The para-spinal
marked degeneration and had been pa
ed, and the presence of nemaline rods

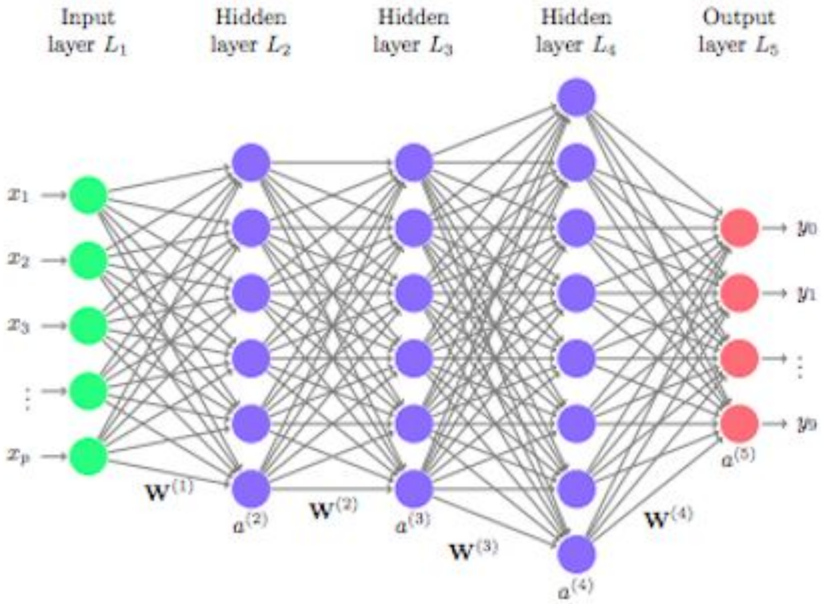
Image credit: nvidia.com, fastdatascience.com

Artificial Intelligence Challenges

Deep Learning, is a type of Machine Learning, inspired by the structure of a human brain.

Deep learning algorithms attempt to draw similar conclusions as humans would by continually analyzing data with a given logical structure. To achieve this, deep learning uses a multi-layered structure of algorithms called neural networks.

Just as we use our brains to identify patterns and classify different types of informations, neural networks can be taught to perform the same tasks on data.



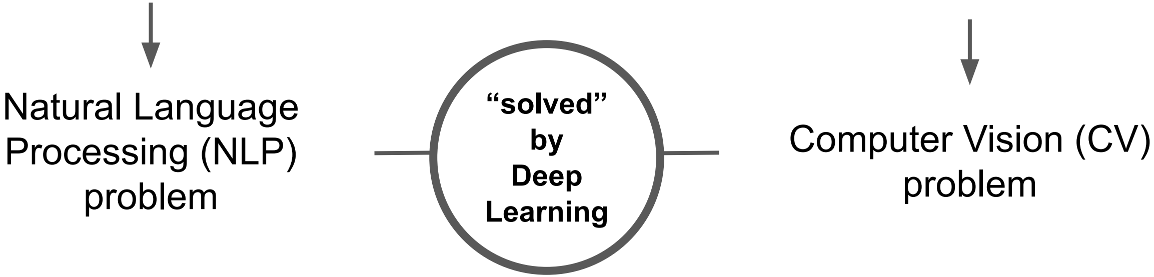
[Sunpark, 2019] - Image credit: UC Business Analytics R Programming Guide

Describing videos as combination of CV and NLP

Commonly, computer vision applications based on deep learning techniques try to predict label from an image, like humans does. But we can do more: we do not only identify an object from an image, but also identify actions into a video or describe what is happening into it.

Describe a video:

"**Create one or more sentences** that explain **what happen in a video**"



Natural Language Processing (NLP) problem

"solved" by Deep Learning

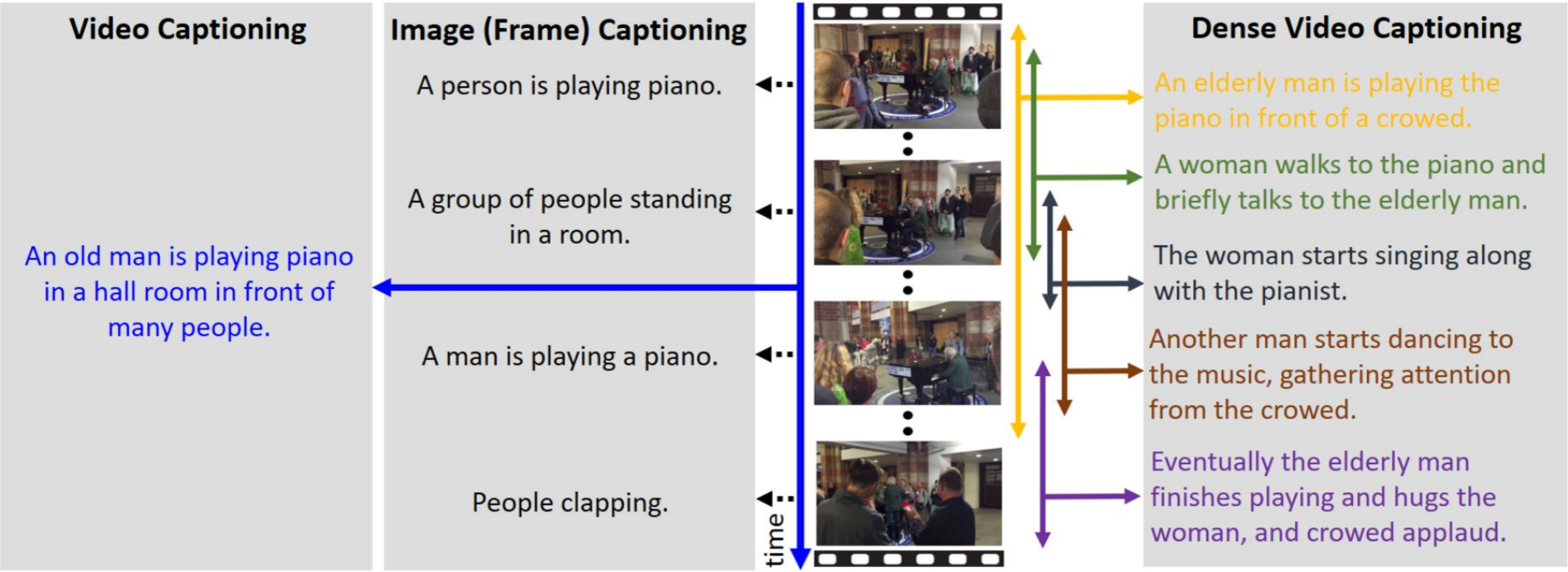
Computer Vision (CV) problem

Before answering "Can a neural networks describe a video?" we should know

"**How we could use a neural network to describe a video?**"

[Islan et al., 2021], [Aafaq et al., 2020]

Video captioning (and friends…)



Video Captioning
An old man is playing piano in a hall room in front of many people.
Image (Frame) Captioning
A person is playing piano.
A group of people standing in a room.
A man is playing a piano.
People clapping.
time
Dense Video Captioning
An elderly man is playing the piano in front of a crowed.
A woman walks to the piano and briefly talks to the elderly man.
The woman starts singing along with the pianist.
Another man starts dancing to the music, gathering attention from the crowed.
Eventually the elderly man finishes playing and hugs the woman, and crowed applaud.

Video captioning is the process where textual descriptions are generated from a sequence of video frames

Image (video frame) **captioning** describes each frame with a single sentence

Dense Video Captioning creates multiple natural language sentences which are used for information enriched, possibly overlapping multiple events of different lengths

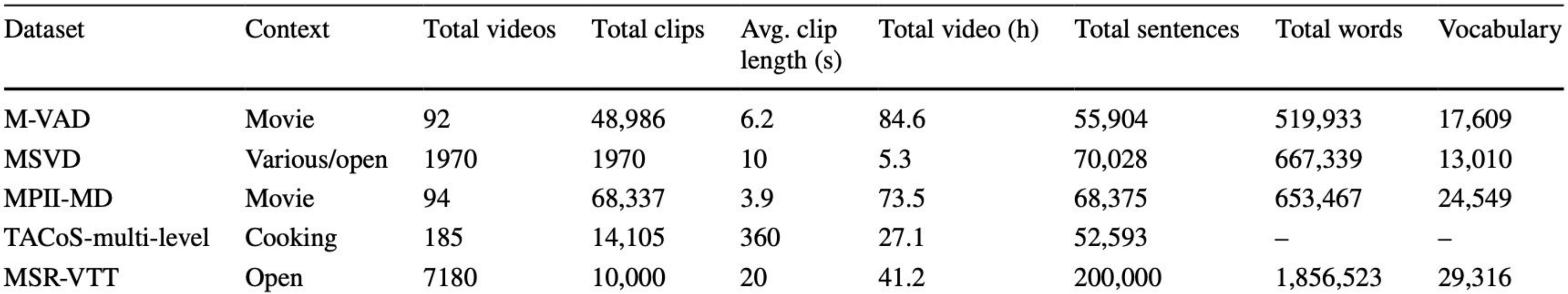
[Islan et al., 2021], [Aafaq et al., 2020]

Benchmark Datasets

Like each ImageNet for image classification task, Video Captioning has own benchmark.

First datasets were created with videos of specific categories like cooking, makeup, movies, etc. These hinder the creation of “general purpose” deep learning models.

Recently new open domain dataset has been realized that made a breakthrough of this filed like MSVD and MSR-VTT.



Dataset	Context	Total videos	Total clips	Avg. clip length (s)	Total video (h)	Total sentences	Total words	Vocabulary
M-VAD	Movie	92	48,986	6.2	84.6	55,904	519,933	17,609
MSVD	Various/open	1970	1970	10	5.3	70,028	667,339	13,010
MPII-MD	Movie	94	68,337	3.9	73.5	68,375	653,467	24,549
TACoS-multi-level	Cooking	185	14,105	360	27.1	52,593	–	–
MSR-VTT	Open	7180	10,000	20	41.2	200,000	1,856,523	29,316

[Islan et al., 2021], [Aafaq et al., 2020]

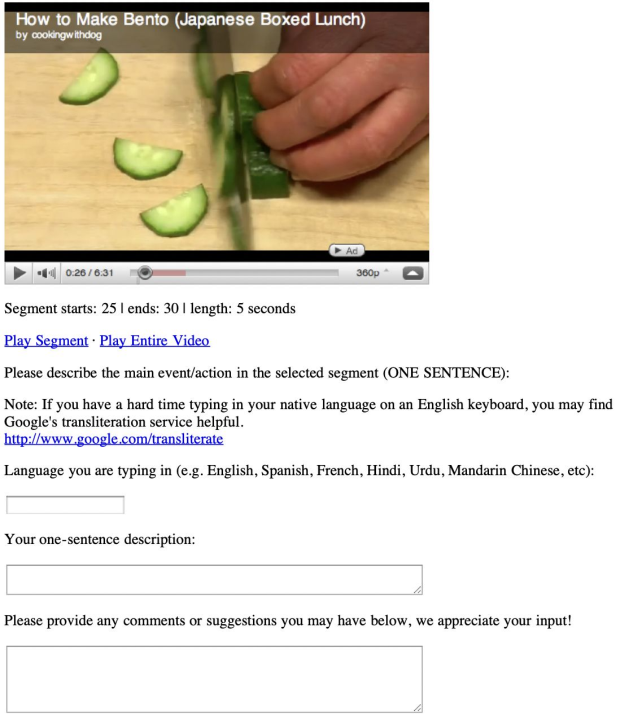
Benchmark Datasets

Microsoft Video Description (MSVD) dataset comprises of 1,970 YouTube clips with human annotated sentences. This dataset was also annotated by AMT (Amazon Mechanical Turk) workers.

The duration of each video in this dataset is typically between 10 to 25 seconds mainly showing one activity.

On average, there are 41 single sentence descriptions per clip.

Almost all research groups have split this dataset into training, validation and testing partitions of 1200, 100 and 670 videos respectively



Segment starts: 25 | ends: 30 | length: 5 seconds

Play Segment · Play Entire Video

Please describe the main event/action in the selected segment (ONE SENTENCE):

Note: If you have a hard time typing in your native language on an English keyboard, you may find Google's transliteration service helpful.
http://www.google.com/transliterate

Language you are typing in (e.g. English, Spanish, French, Hindi, Urdu, Mandarin Chinese, etc):

Your one-sentence description:

Please provide any comments or suggestions you may have below, we appreciate your input!

[Aafaq et al., 2020], [Chen et al., 2011]

Benchmark Datasets

MSR-Video to Text (MSR-VTT) is one of the largest video captioning open domain datasets. It comprises of 7180 videos subdivided into 10,000 clips. The clips are grouped into 20 different categories.

The duration of each clip is between 10 and 30 seconds.

The dataset is splitted into 6513 training, 497 validation and 2990 test videos. Each video comprises 20 reference captions annotated by AMT (Amazon Mechanical Turk) workers.

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.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical
3. Dancers are playing a routine.
4. People are dancing in a musical.
5. Some people are acting and singing for performance.

1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

[Aafaq et al., 2020], [Xu et al., 2016]

How to evaluate a Video Captioning model?

Most of the popular benchmark datasets give multiple sentences that describe a video. This because there are multiple way to interpret a observation. These sentences that may differ not only syntactically but also in terms of semantic content.

Video Captioning models generate only one sentence (**candidate**) that we can compare with multiple **reference** sentences.

How we can evaluate video descriptions?

There is no specific metric dedicated to the video captioning task. Studies commonly use metric borrowed from other task like machine translation or image captioning.

DEVELOPING NOW
HILLARY CLINTON SPEAKS ABOUT CHARLESTON MASSACRE
CALLED FOR GUN REFORM DAY AFTER SHOOTING
msnbc

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

} References

A lady is making a speech on new channel. } Candidate

[Islan et al., 2021], [Aafaq et al., 2020]

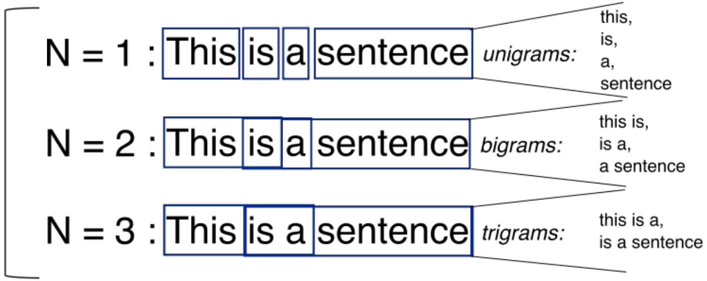
How to evaluate a Video Captioning model?

Sentences are decomposed in n-grams.

Then n-grams are used to calculate Precision and Recall.

Precision: number of candidate n-grams occurring in any reference divided by total number of n-grams in the candidate sentence.

Recall: number of candidate n-grams occurring in any reference divided by total number of n-grams in references sentences.



Example 1

Candidate 1: the the the the the the the
Reference: the cat is on the mat

The uni-gram precision of candidate 1 is 2/7 (28.5%)
The uni-gram recall of candidate 1 is 2/6 (33.3%)

Example 2

Candidate 2: the cat is mat the on
Reference: the cat is on the mat

The uni-gram precision of candidate 2 is 6/6 (100%)
The uni-gram recall of candidate 2 is 6/6 (100%)

[Kilickaya et al., 2016], [Papineni et al., 2002]

Metrics

Metric Name	Designed for	Methodology
BLEU	Machine translation	n-gram precision
ROUGE	Document summarization	n-gram recall
METEOR	Machine translation	n-gram with synonym matching
CIDEr	Image captioning	tf-idf weighted n-gram similarity

BLEU scores take into account the overlap between predicted uni-grams (single word) or higher order n-gram (sequence of n adjacent words) and a set of one or more reference sentences. The more the number of matches between candidate and reference translation, the better is the machine translation (max value is 1).

ROUGE evaluation is done via comparing overlapping n-grams, word sequences and word pairs. In Video Captioning the ROUGE-L version is typically used, which basically measures the longest common subsequences between a pair of sentences.

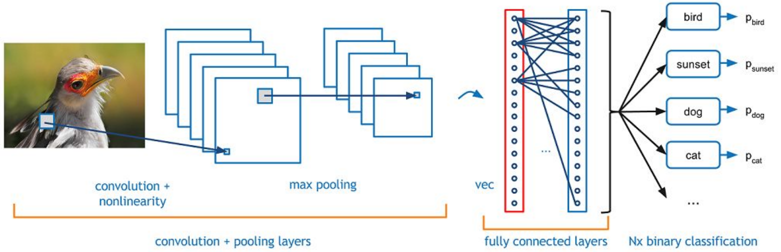
METEOR is defined as the harmonic mean of precision and recall of n-gram matches between sentences. Additionally, it makes use of WordNet-based synonym matching.

For calculating **CIDEr** metric, an initial stemming is applied and each sentence is represented with a set of n-grams. Then, the co-occurrences of n-grams in the reference sentences and candidate sentence are calculated. The n-grams are down-weighted using tf-idf. Finally, the cosine similarity between n-grams of the candidate and the references is computed.

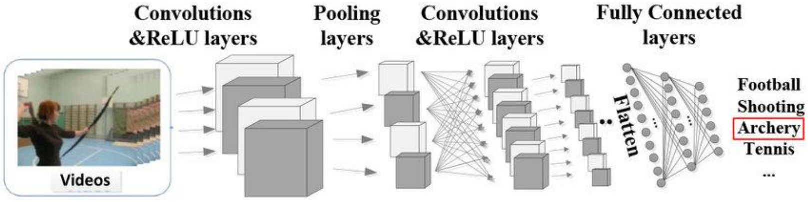
[Kilickaya et al., 2016]

Baseline model

Understand video frames

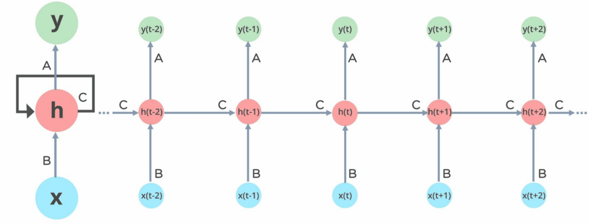


Pre-trained 2D Convolutional Neural Network on image classification task

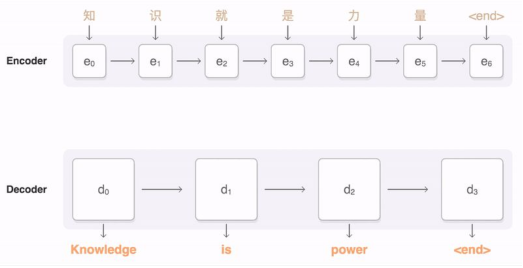


Pre-trained 3D Convolutional Neural Network on action/video classification task

Generate sentence



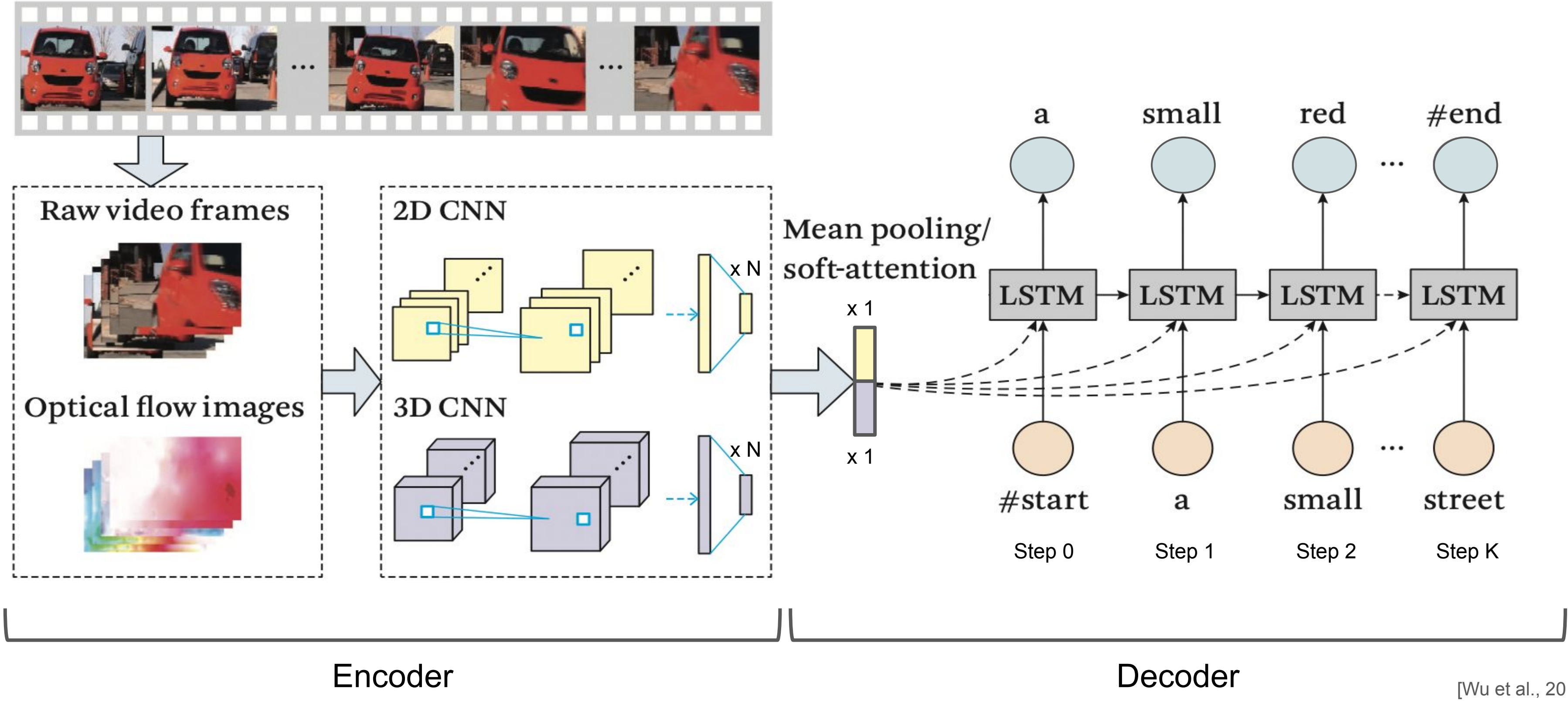
Recurrent Neural Network (usually LSTM or GRU)



Attention Mechanism

Image credit: dshaid380, Junzhong Shen, Avijeet Biswal, Google

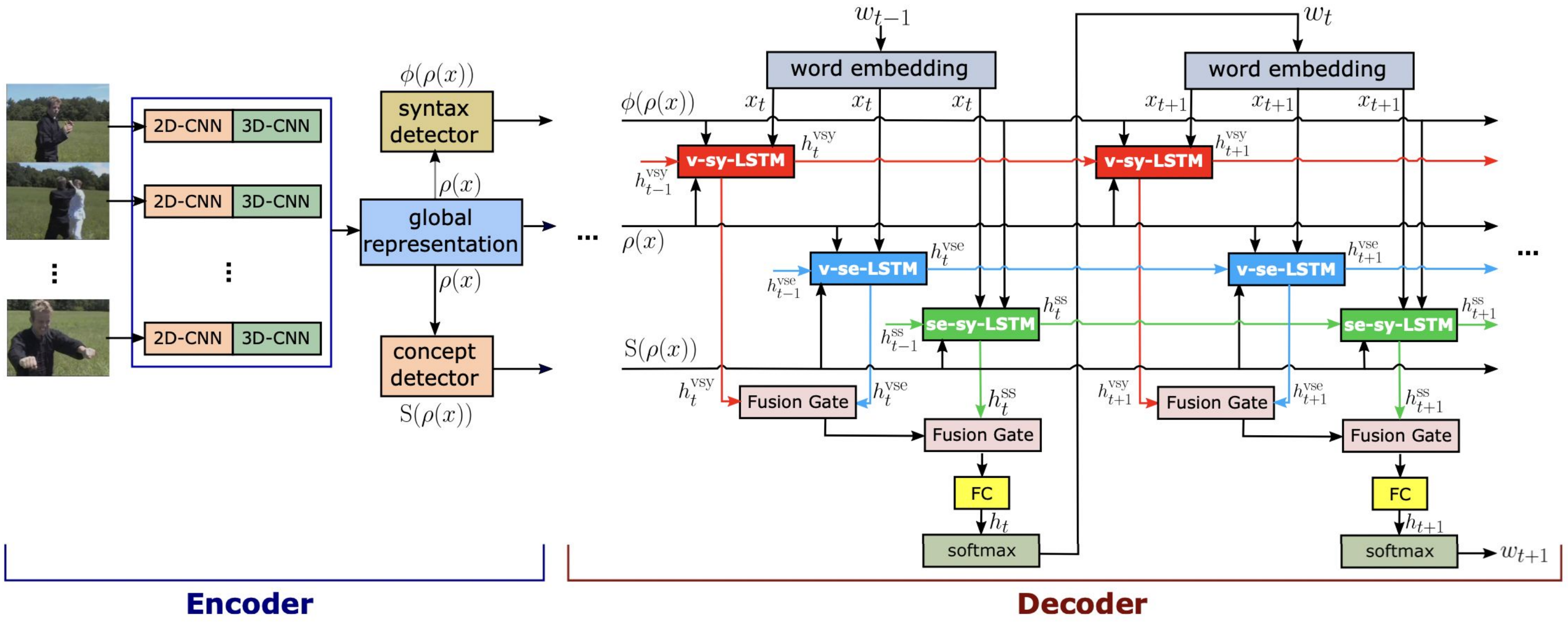
Baseline model



[Wu et al., 2018]

State of the art model

Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding (SemSynAN)



[Perez-Martin et al., 2021]

State of the art model

Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding (SemSynAN)

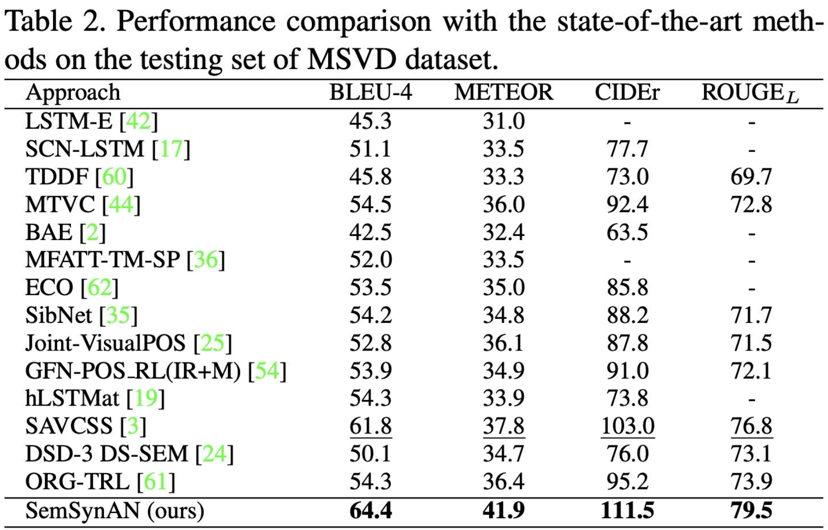


Table 2. Performance comparison with the state-of-the-art methods on the testing set of MSVD dataset.

Approach	BLEU-4	METEOR	CIDEr	ROUGE$_L$
LSTM-E [42]	45.3	31.0	-	-
SCN-LSTM [17]	51.1	33.5	77.7	-
TDDF [60]	45.8	33.3	73.0	69.7
MTVC [44]	54.5	36.0	92.4	72.8
BAE [2]	42.5	32.4	63.5	-
MFATT-TM-SP [36]	52.0	33.5	-	-
ECO [62]	53.5	35.0	85.8	-
SibNet [35]	54.2	34.8	88.2	71.7
Joint-VisualPOS [25]	52.8	36.1	87.8	71.5
GFN-POS_RL(IR+M) [54]	53.9	34.9	91.0	72.1
hLSTMat [19]	54.3	33.9	73.8	-
SAVCSS [3]	61.8	37.8	103.0	76.8
DSD-3 DS-SEM [24]	50.1	34.7	76.0	73.1
ORG-TRL [61]	54.3	36.4	95.2	73.9
SemSynAN (ours)	**64.4**	**41.9**	**111.5**	**79.5**

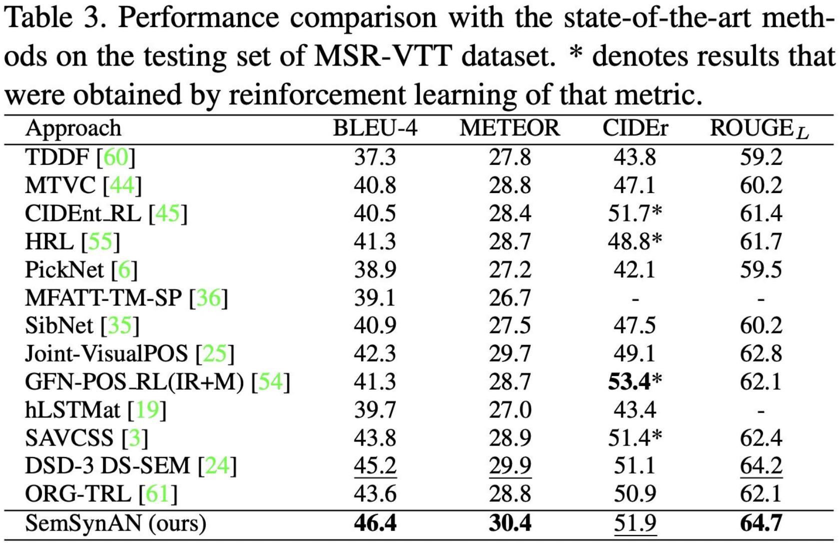


Table 3. Performance comparison with the state-of-the-art methods on the testing set of MSR-VTT dataset. * denotes results that were obtained by reinforcement learning of that metric.

Approach	BLEU-4	METEOR	CIDEr	ROUGE$_L$
TDDF [60]	37.3	27.8	43.8	59.2
MTVC [44]	40.8	28.8	47.1	60.2
CIDEnt_RL [45]	40.5	28.4	51.7*	61.4
HRL [55]	41.3	28.7	48.8*	61.7
PickNet [6]	38.9	27.2	42.1	59.5
MFATT-TM-SP [36]	39.1	26.7	-	-
SibNet [35]	40.9	27.5	47.5	60.2
Joint-VisualPOS [25]	42.3	29.7	49.1	62.8
GFN-POS_RL(IR+M) [54]	41.3	28.7	**53.4***	62.1
hLSTMat [19]	39.7	27.0	43.4	-
SAVCSS [3]	43.8	28.9	51.4*	62.4
DSD-3 DS-SEM [24]	45.2	29.9	51.1	64.2
ORG-TRL [61]	43.6	28.8	50.9	62.1
SemSynAN (ours)	**46.4**	**30.4**	51.9	**64.7**

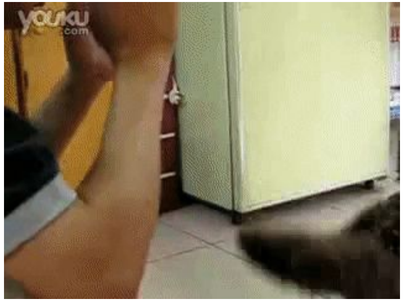
[Perez-Martin et al., 2021]

State of the art model

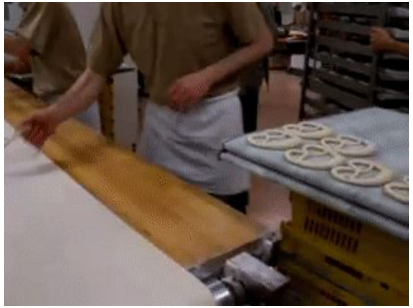
✓ a man is riding on a motorcycle

✓ a cat is cleaning its face

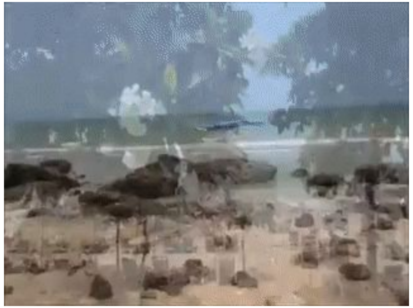
✓ a person is folding a paper airplane



✗ a monkey is playing with a dog



✗ a men is a pizza



✗ a person is recording the beautiful beautiful beautiful beautiful

[Perez-Martin et al., 2021]

Conclusions

Video captioning is one of the biggest challenges for Artificial Intelligence. To achieve these results we need to create a model that is capable to understand video frame (computer vision) and generate sentences for it (natural language processing).

“**How we could use a neural network to describe a video?**”

Deep neural networks allows to create hight customized models, so we can definitely create good video captioning solutions for it. Today the main common solution is to use the encoder-decoder paradigm.

“**Can a neural networks describe a video?**”

Yes and no… the results achieved so far do not solve this task at 100%. Right now state-of-the-art solutions lack of acceptable results in order to be used in business activities, Video captioning is still an open problem. As usual, one big issue is benchmark datasets that contain few examples.

References

Video captioning studies

[Islan et al., 2021] Exploring Video Captioning Techniques: A Comprehensive Survey on Deep Learning Methods

[Perez-Martin et al., 2021] Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding

[Aafaq et al., 2020] Video Description: A Survey of Methods, Datasets and Evaluation Metrics

[Wu et al., 2018] Deep Learning for Video Classification and Captioning

Metrics

[Kilickaya et al., 2016] Re-evaluating Automatic Metrics for Image Captioning

[Papineni et al., 2002] BLEU: a Method for Automatic Evaluation of Machine Translation

Benchmark Dataset

[Xu et al., 2016] MSR-VTT: A Large Video Description Dataset for Bridging Video and Language

[Chen et al., 2011] Collecting Highly Parallel Data for Paraphrase Evaluation

Articles

[Sunpark, 2019] It’s Deep Learning Times: A New Frontier of Data

Thanks ;)

Silvio Olivastri - Senior Data Scientist

{codemotion}

Online Tech Conference

- Italian edition -

23-24-25 Marzo, 2021