Video Captioning (to be reviewed)

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

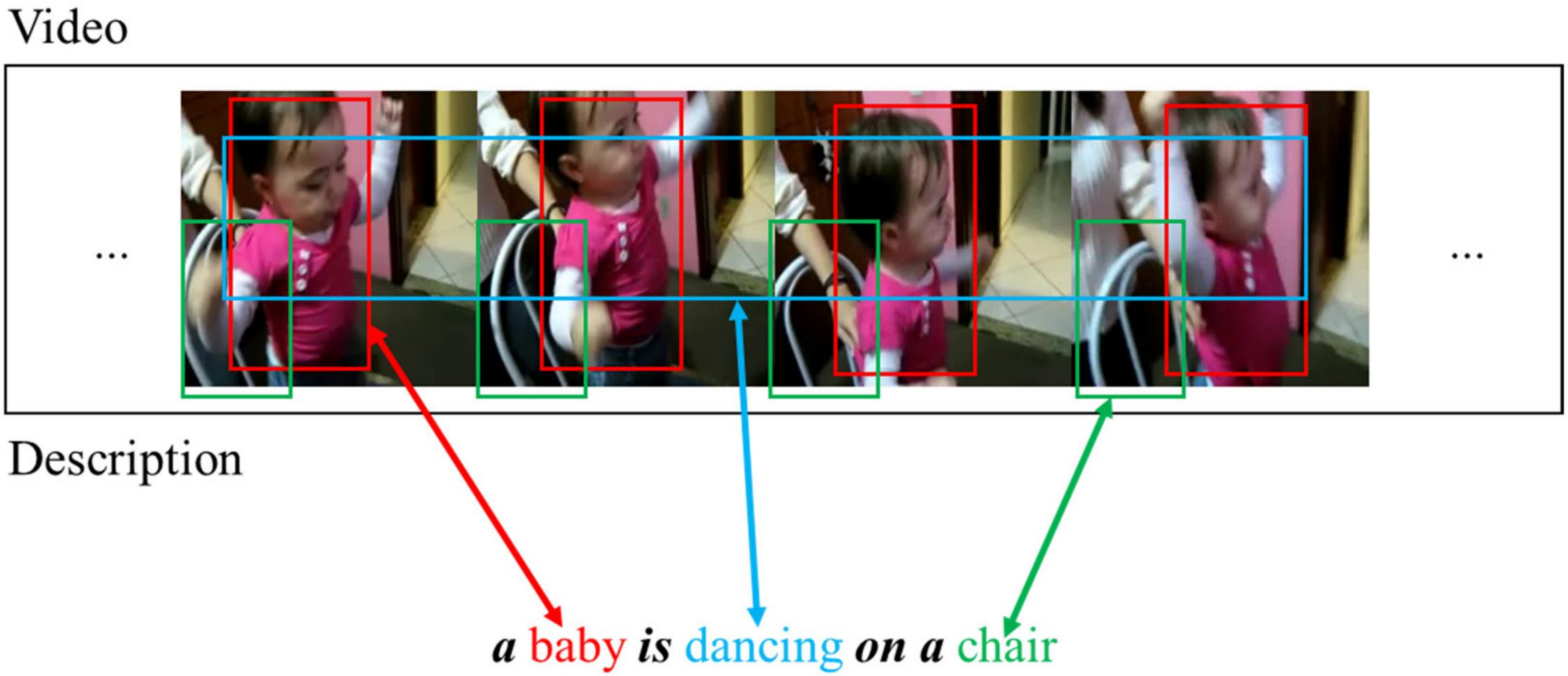


Image from mdpi

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

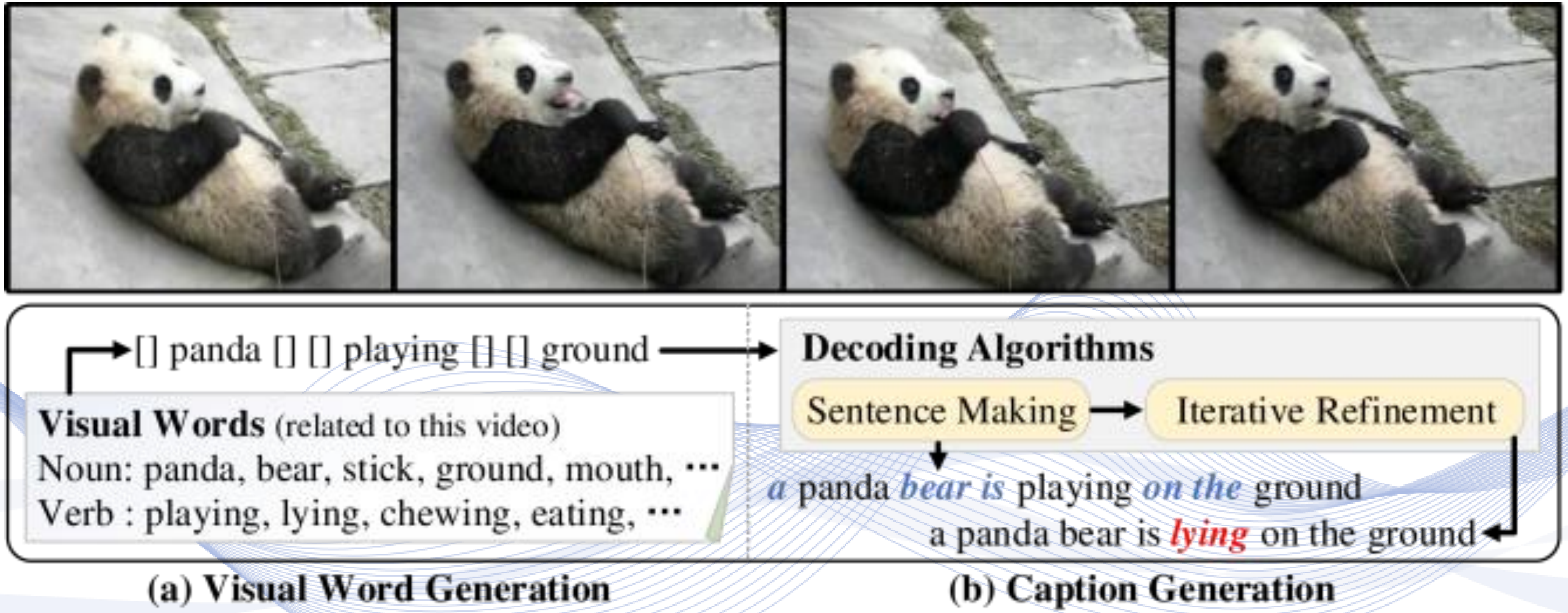


Image from arxiv-vanity

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

Caption:

A caption is the title of a magazine article, a descriptive title under a photograph, the words at bottom of a television or movie screen to translate the dialogue into another language or to provide the dialogue to the hard of hearing. A caption generally, may be a few words or several sentences. [HTT2020]

Video Captioning

Caption:

Captions for a Image are the little “headlines” over the “cutlines” (the words describing the photograph).[HTT2013]

Image Caption (Image from https://xiangliu.ca/image-caption/)

Video Captioning

The importance of captioning lies in its ability to make video more accessible in numerous ways.

It allows d/Deaf and hard of hearing individuals to watch videos, helps people to focus on and remember the information more easily, and lets people watch it in sound-sensitive environments. [LED2018]

Video Captioning

Unlike image captioning, which aims to describe a static scene, video captioning is a more provocative sense that a series of coherent scenes need to be understood in order to create multiple sections of description together. [HTT2019]

Video Captioning

Video captions are similar to a transcription, but are synced to a video’s time codes, allowing the viewer to follow along with a video’s words as they’re being said. They ‘re also shown within the video player in a seamless and unobtrusive way. [HTT2019]

Video Captioning

Video captioning is process of summarizing the content, event and action of the video into a short textual form which can be helpful in many research areas such as video guided machine translation, video sentiment analysis and providing aid to needy individual. [THO2020]

Video Captioning

Video Captioning is one of the kind of **Static video summarization** and it generates a textual description for a given video content.

Video captioning problem arises naturally as the very next step where a sentence is generated to describe a video clip that captures its visual semantics. [PRA2013]

Video Captioning

It is a task of automatic captioning a video by understanding the action and event in the video which can help in the retrieval of the video efficiently through text. On addressing the task of video captioning effectively, the gap between computer vision and natural language can also be minimized. [ZAC2012]

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

The **template-based** approaches use predefined templates for generating the captions by fitting the attributes identified in the video.

These kinds of approaches need the proper alignment between the words generated for the video and the predefined templates. [JIA2018]

Video Captioning

In contrast to template-based approach, the **sequence learning** based approach learn the sequence of word conditioned on previously generated word and visual feature vector of the video.

This approach is commonly used in Machine Translation (MT) where the target language (T) is conditioned on the source language (S). [YAN2016]

Video Captioning

Video Captioning (Image from ResearchGate

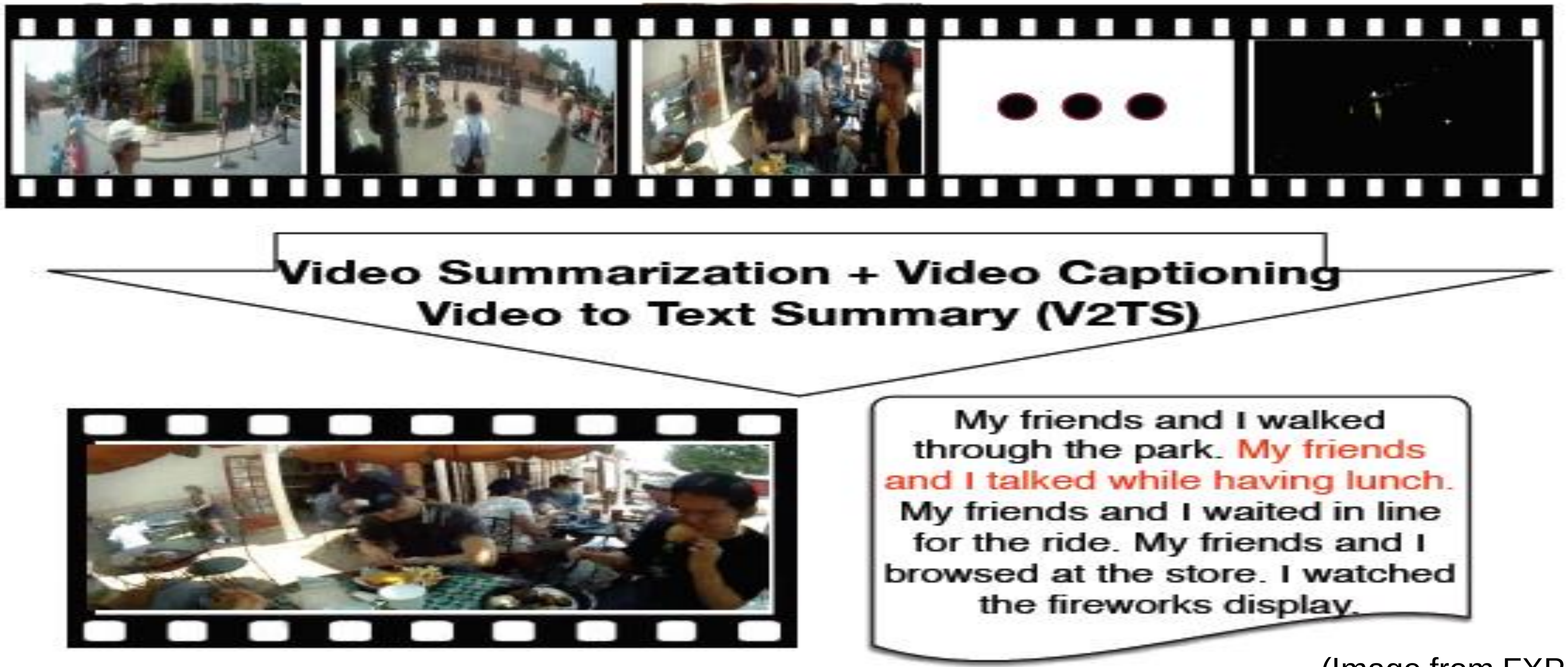
Artificial Intelligence & Information Analysis Lab

Video Captioning

The video captioning is the quite challenging topic because of the complex and diverse nature of video content.

However, the understanding between video content and natural language sentence remains an open problem to create several methodology to better understand the video and generate the sentence automatically. [HUA2021]

Video Captioning



(Image from FXPAL)

Artificial Intelligence & Information Analysis Lab

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

- **Types**
  - **Closed Captions**

These are hidden on the 21st line of the vertical blanking interval (VBI) of a video signal and are made visible by a decoder at the time of viewing. They are usually white letters encased in a black box. [HTT2021]

Video Captioning: Captioning Types, Methods, and Styles

Closed Captions

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

Video Captioning: Captioning Types, Methods, and Styles

- **Types**
  - **Subtitles**

Subtitles are usually white or yellow letters with a black rim or drop shadow. Some are always visible, like the ”open captions” of DCMP videos. Others, like those on DVD and the Internet, are displayed utilizing the medium’s menu option. [HTT2021]

Video Captioning: Captioning Types, Methods, and Styles

Subtitles

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

Video Captioning: Captioning Types, Methods, and Styles

- **Types**
  - **Subtitles for the Deaf and Hard of Hearing (SDH)**

These are just like subtitles, but SDH includes information such as sound effects, speaker identification, and other essential nonspeech features. These are presented as close to verbatim as possible. [HTT2021]

Video Captioning: Captioning Types, Methods, and Styles

- **Types**
  - **Subtitles for the Deaf and Hard of Hearing (SDH)**

Foreign Film Subtitles, which are written for hearing viewers, usually do not indicate information other than dialogue, and often are edited. Some may translate important onscreen printed information such as a street sign or a written message. [HTT2021]

Artificial Intelligence & Information Analysis Lab

Video Captioning: Captioning Types, Methods, and Styles

Subtitles for the Deaf and Hard of Hearing (SDH)

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: Captioning Types, Methods, and Styles

- **Methods**
  - **Off-line**

Off-line captions are created and added after a video segment has been recorded and before it is aired or played. Examples of programs that utilize o-line captioning are prime-time TV programs, made-for-TV movies, and educational media. [HTT2021]

Video Captioning: Captioning Types, Methods, and Styles

- **Methods**
  - **On-line**

    On-line captions are created and displayed at the time of program origination, and sometimes referred to as Real-time. Examples of programming that utilizes on-line captioning are sporting events, newscasts, and other events that do not allow time to prepare off-line captions. [HTT2021]

Video Captioning: Captioning Types, Methods, and Styles

- **Styles**
  - **Roll-up**

Roll-up captions are usually verbatim and synchronized. Captions follow double chevrons (which look like ”greater than” symbols), and are used to indicate different speaker identifications. Each sentence ”rolls up” to about three lines. The top line of the three disappears as a new bottom line is added, allowing the continuous rolling up of new lines of captions. [HTT2021]

Video Captioning: Captioning Types, Methods, and Styles

- **Styles**
  - **Paint-on**

Paint-on captions are very similar to roll-up captions. Individual words are ”painted on” from left to right, not popped on with all captions at once, and usually are verbatim. [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

A good video captioning requires both local and global understanding, recognizing activities and reasoning dependencies between local activities and context.

Each subsection below focuses on one methodology of approaching video captioning problem, and discusses both the backbone and various variants of it as well as its advantages and limitations, from classical ones to state-of-the-art ones. [JIA2018]

Methodology of approaching Video Captioning problem

Template-based Captioning

Following the success of image recognition and activity recognition, one naive approach is to synthesize the detected outputs into a sentence using a template to ensure grammatical correctness. [JIA2018]

Methodology of approaching Video Captioning problem

Template-based Captioning

Template-based language methods first split sentences into fragments (e.g. subject, verb and object) following specific rules of language grammar, and each fragment is associated with detected words (e.g. objects, actions and attributes) from visual content. Then generated fragments are composed to a sentence with predefined language template. [JIA2018]

Methodology of approaching Video Captioning problem

Template-based Captioning

As a result, the captioning quality highly depends on the templates of sentence and sentences are always generated with syntactical structure.

Although template-based language can generate complete sentences, generated descriptions are very rigid. [JIA2018]

Methodology of approaching Video Captioning problem

Template-based Captioning

Meanwhile, the evaluation is usually limited to narrow domain with a small vocabulary, such as TACoS dataset. For any sufficiently rich domain, the required complexity of rules and templates makes manual design of templates unfeasible or too expensive. [JIA2018]

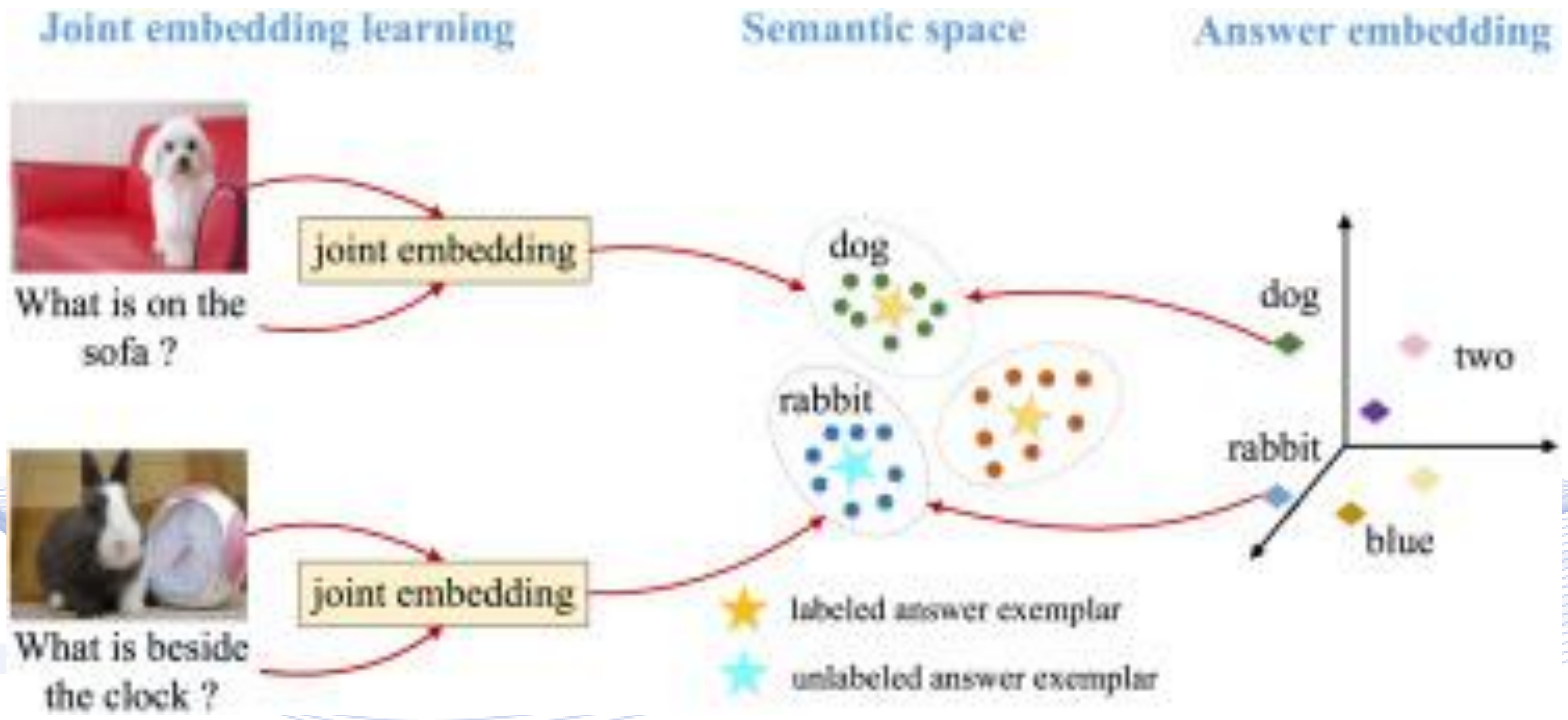
Methodology of approaching Video Captioning problem

Joint Embedding

Video captioning problem arises as a side product of video retrieval problem where a video is to be retrieved according to given text description. Since multi-model embedding is a common practice to solve video retrieval problem, some early works apply joint embedding approach to video and language for video captioning as well. [JIA2018]

Methodology of approaching Video Captioning problem

Joint Embedding



(Image from ScienceDirect)

Methodology of approaching Video Captioning problem

Joint Embedding

The framework of joint embedding consists of **three components**: **(1)** a visual model to map video to representation vector, **(2)** a language model to map text caption to representation vector, **(3)** a projection of visual representation vector and language representation vector to the shared space, by minimizing distance between the two projected vectors. [JIA2018]

Methodology of approaching Video Captioning problem

Joint Embedding

The idea is that the joint embedding space is semantically continuous and ensures semantically similar items, regardless of being video or description, are close to each other. During inference time, an input video is mapped to a point in the shared space corresponding to a semantically close sentence description which is further converted to text in the inverse process of the language model. [JIA2018]

Methodology of approaching Video Captioning problem

Joint Embedding

There are many possible choices of visual model and language model as practiced in the literature, such as:

The **simplistic** form of language model could be taking bag of words or one-hot encoding as semantic representation. Based on the assumption that essential semantic meaning of a video can be captured by SVO (Subject, Verb, Object) triplets. [JIA2018]

Methodology of approaching Video Captioning problem

Joint Embedding

The **visual** model follows the progress of deep models in image domain.

In general, the approach of joint embedding is effective in the scenario of videos within narrow domain since the embedding space can generalizes such finite domain well, and richer model structures boost up performance. [JIA2018]

Methodology of approaching Video Captioning problem

Joint Embedding

However, it can easily fail when encountering videos with situations that haven’t been seen before. Also since the embedding is of fixed length, it limits the amount of information that can be carried by video and text description. [JIA2018]

Methodology of approaching Video Captioning problem

Encoder-Decoder

Inspired by the progress in machine translation and image captioning, some other early works formulate video captioning problem partially as machine translation problem where a semantic representation is generated for a video and then is translated to natural language sentence. [JIA2018]

Methodology of approaching Video Captioning problem

Encoder-Decoder

The framework those works propose is an Encoder-Decoder structure that encodes video into semantic representation and then decodes into natural language. The benefit of translation is that now we can have an open world vocabulary if we feed machine translation model with large text corpus, which is not hard to obtain. [JIA2018]

Methodology of approaching Video Captioning problem

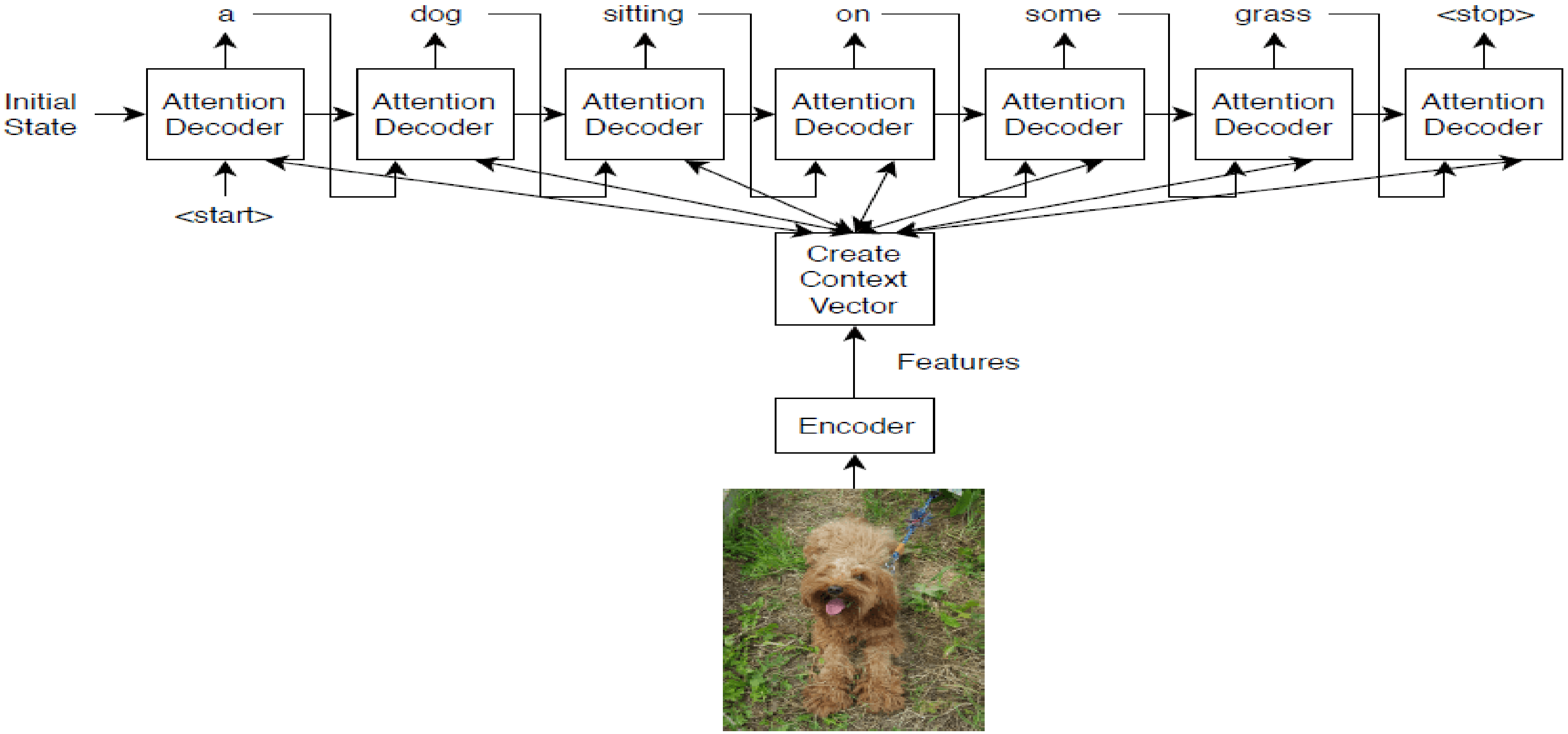
Encoder-Decoder Mechanisms:

- **Attention Mechanism**

The attention mechanisms in deep neural networks are inspired by human's attention that sequentially focuses on the most relevant parts of the information over time to make predictions. [JIA2018]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms: Attention Mechanism



(Image from MathWorks)

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Attention Mechanism**

The recently proposed soft attention mechanism to balance exploitation of local temporal structure, which captures details of activities, and global temporal structure, which reflects long-term dependencies and ordering of activities. [JIA2018]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Attention Mechanism**

The framework first uses 3D-CNN to generate temporal features vectors which capture local temporal structure (motion features). The decoder is an LSTM with soft attention mechanism, which takes in the dynamic weighted sum of the temporal feature vectors according to attention weights generated at each time step. [JIA2018]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Attention Mechanism**

Specifically, attention weights are generated for all the frames based on hidden state of previous time step (which presumably summarizes all the previously generated words) and the corresponding frame's temporal feature vector. [JIA2018]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Attention Mechanism**

Soft attention mechanism enables the decoder to look at different temporal locations and relate activities occurring cross time span for global reasoning. It has become a common practice in future works. [JIA2018]

Methodology of approaching Video Captioning problem

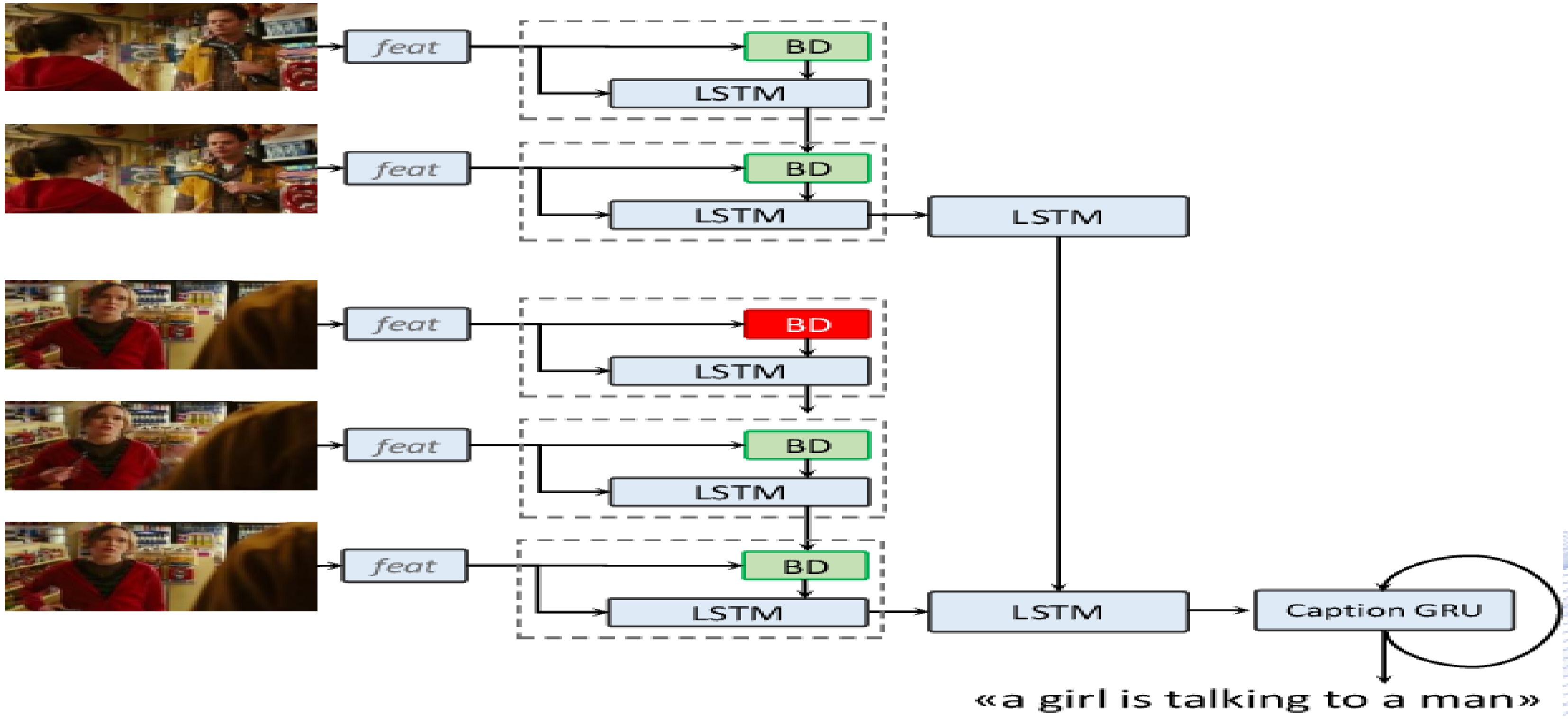
Encoder-Decoder Mechanisms:

- **Hierarchical Neural Encoder**

Another line of works focuses on refining neural encoder. Even though LSTM can deal with long video clips in principle, it has been reported that the favorable length of video clips to LSTM falls in the range of 30 to 80 frames. [JIA2018], [MOO2015]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms: Hierarchical Neural Encoder



(Image from SemanticScolar)

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Hierarchical Neural Encoder**

Therefore, it’s usually hard fora plain LSTM to capture the large number of long-range dependencies in video. Aiming at learning the visual features with multiple temporal granularities, are used Hierarchical Recurrent Neural Encoder (HRNE). [JIA2018], [MOO2015]

Artificial Intelligence & Information Analysis Lab

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Hierarchical Neural Encoder**

Hierarchical Recurrent Neural Encoder (HRNE), consists of a LSTM filter on subsequences of an input sequence to explore local temporal features within sub-sequences and then another layer of LSTM on top to summarize and learn temporal dependencies among subsequences. [JIA2018], [MOO2015]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Hierarchical Neural Encoder**

Such a hierarchical structure significantly reduces the length of input information follow but is still capable of exploiting temporal information over longer time. It has been noted that more LSTM layers could be added to HRNE to build multiple time-scale abstraction of the visual information. [JIA2018], [MOO2015]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Hierarchical Neural Encoder**

The method achieves state-of-the-art performance on video captioning benchmarks at that time. However, it requires fixed manual setting of the sub-sequence length, and thus it doesn’t adapt to varying types of videos. [JIA2018], [MOO2015]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Paragraph Description**

This line of works focuses on generating a long story-like caption. Some works first temporally segment the video with action localization or different levels of details, and then generate multiple captions for those segments and connect them with natural language processing techniques. [JIA2018], [SHI2016], [QIU2014]

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:

- **Paragraph Description**

The key framework proposed by "Video paragraph captioning using hierarchical recurrent neural networks" is hierarchical RNN (h-RNN) for describing a long video with a paragraph consisting of multiple sentences. This framework consists of two generators:[JIA2018], [HUA2016]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Paragraph Description**

(1) a sentence generator which produces single short sentences that describe specific time intervals and video regions, and **(2)** a paragraph generator which takes the sentential embedding as input and uses another recurrent layer to output the paragraph state; such state is then used to initialize the sentence generator. [JIA2018], [HUA2016]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Paragraph Description**

In addition, both sentence and paragraph generators adopt recurrent layers for language modeling. It uses C3D features to model video motion and activities, and applies soft temporal attention to the feature pool before feeding into Hierarchical RNN. number of sentences in the paragraph is 1. [JIA2018], [HUA2016]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Paragraph Description**

The model is evaluated on TACoS-Multi Dataset which provides paragraph description to video clips and MSVD which provides parallel sentences to video clip and is used as a special case where the number of sentences in the paragraph is 1. [JIA2018], [HUA2016]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Paragraph Description**

Interestingly, the experiments show that the special case hRNN outperforms state-of-the-art single-sentence captioning methods on MSVD dataset at that time, which means the hierarchy helps not only inter-sentence dependencies but also intra-sentence dependencies. [JIA2018], [HUA2016]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Paragraph Description**

Meanwhile, h-RNN definitely outperforms baseline methods that have no hierarchy, i.e., with only the sentence generator, but not the paragraph generator. [JIA2018], [HUA2016]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Paragraph Description**

The evaluation of paragraph generation has only been conducted on closed-domain dataset, and thus the conclusion is not necessarily applicable to general open domain dataset. This calls for large-scale open domain video dataset with paragraph description annotations. [JIA2018], [HUA2016]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

The dense captioning task generalizes object detection when the descriptions consist of a single word, and Image Captioning when one predicted region covers the full image. [JIA2018], [REN2017]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

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

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

Dense-captioning events in a video involves detecting multiple events that occur in a video and describing each event using natural language. These events are temporally localized in the video with independent start and end times, resulting in some events that might also occur concurrently and overlap in time. [JIA2018], [REN2017]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

Such models would likely concentrate on an elderly man playing the piano in front of a crowd. While this caption provides us more details about who is playing the piano and mentions an audience, it fails to recognize and articulate all the other events in the video. [JIA2018], [REN2017]

Artificial Intelligence & Information Analysis Lab

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

For example, at some point in the video, a woman starts singing along with the pianist and then later another man starts dancing to the music. [JIA2018], [REN2017]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

In order to identify all the events in a video and describe them in natural language, we introduce the task of dense-captioning events, which requires a model to generate a set of descriptions for multiple events occurring in the video and localize them in time. [JIA2018], [REN2017]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

Dense-captioning events is analogous to dense-image-captioning; it describes videos and localize events in time whereas dense-image-captioning describes and localizes regions in space. [JIA2018], [REN2017]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

However, we observe that dense-captioning events comes with its own set of challenges distinct from the image case. One observation is that events in videos can range across multiple time scales and can even overlap. [JIA2018], [REN2017]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

While piano recitals might last for the entire duration of a long video, the applause takes place in a couple of seconds.

To capture all such events, we need to design ways of encoding short as well as long sequences of video frames to propose events. [JIA2018], [REN2017]

Methodology of approaching Video Captioning problem

Encoder-Decoder Mechanisms:

- **Dense Captioning**

Past captioning works have circumvented this problem by encoding the entire video sequence by mean-pooling or by using a recurrent neural network (RNN). While this works well for short clips, encoding long video sequences that span minutes leads to vanishing gradients, preventing successful training. [JIA2018], [REN2017]

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-SVO

The **SVO** (Subject, Object, Verb) tuples based methods are among the first successful methods used specifically for video description.

However, research efforts were made long before to describe visual content into natural language, albeit not explicitly for captioning or description. [DON2014], [LIU2019]

Artificial Intelligence & Information Analysis Lab

Video Captioning: Evaluation Metrics -SVO

Numerous method have been proposed for detecting objects, humans, actions, and events in videos. [DON2014], [LIU2019]

- **Object Recognition**
- **Human and Activity Detection**
- **Integrated Approaches**

Video Captioning: Evaluation Metrics -SVO

Object Recognition:

Object recognition in SVO approaches was performed typically using conventional methods, including model-based shape matching through edge detection or color matching, HAAR features matching, context-based object recognition, Scale Invariant Feature Transform (SIFT), discriminatively trained part based models and Deformable Parts Model (DPM). [DON2014], [LIU2019]

Video Captioning: Evaluation Metrics -SVO

Human and Activity Detection:

Human detection methods employed features such as Histograms of Oriented Gradient (HOG) followed by SVM. For activity detection, features like Spatiotemporal Interest Points such as Histogram of Oriented Optical Flow (HOOF), Bayesian Networks (BN), Dynamic Bayesian Networks (DBNs), Hidden Markov Models (HMM), state machines, and PNF Networks have been used by SVO approaches. [DON2014], [LIU2019]

Video Captioning: Evaluation Metrics -SVO

Integrated Approaches:

Instead of detecting the description-relevant entities separately, Stochastic Attribute Image Grammar (SAIG) and Stochastic Context Free Grammars (SCFG), allow for compositional representation of visual entities present in a video, an image or a scene based on their spatial and functional relations. Using the visual grammar, the content of an image is first extracted as a parse graph. [DON2014], [LIU2019]

Video Captioning: Evaluation Metrics -SVO

Integrated Approaches:

A parsing algorithm is then used to find the best scoring entities that describe the video.

In other words, not all entities present in a video are of equal relevance, which is a distinct feature of this class of methods compared to the aforementioned approaches. [DON2014], [LIU2019]

Video Captioning: Evaluation Metrics -BLEU

BLEU is one of the most popular metrics in the eld 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 - BLEU

BLEU is calculated as,

$$logBLEU = \min(1 - \frac{l_r}{l_c}, 0 + \sum_{n=1}^{N} w_n \ \ logp_n$$

Artificial Intelligence & Information Analysis Lab

Video Captioning: Evaluation Metrics - BLEU

l_r/l_c : The ratio between the lengths of the corresponding reference corpus and the candidate description,

w_n: The positive weights,

p_n:The geometric average of the modified n-gram precisions.

The second term computes the actual match score,

The first term is a brevity penalty that penalizes descriptions that are shorter than the reference description.

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]

Artificial Intelligence & Information Analysis Lab

Video Captioning: Evaluation Metrics- ROUGE-N

ROUGE-N is computed as,

$$ROUGE-N = \frac{\sum_S \in R_{Sum} \sum_g \in s\, C_m\,(g_n)}{\sum_S \in R_{Sum} \sum_g \in s\, C\,(g_n)}$$

n: The n-gram length,

g_n, and $C_m\,(g_n)$: the highest number of n-grams that are present in candidate as well as ground truth summaries and R_{sum} : Reference summaries. [PAR2017]

Artificial Intelligence & Information Analysis Lab

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 - CIDER

$\boldsymbol{CIDEr_n}$score is computed as,

$$CINDER_n\,(C_i, S_i\,) = \frac{1}{m}\sum_J \frac{g^n(C_i).\,g^n(S_{ij})}{\|g^n(C_i)\|.\,\|g^n(S_{ij})\|}$$

g^n:A vector representing all n–grams with length n and

$g^n(C_i)$: The magnitude of $g^n(C_i)$.

Same is true for $g^n(S_{ij})$. [PAR2017]

Video Captioning: Evaluation Metrics - CIDER

Further, CIDEr uses higher order n-grams (higher the order, longer the sequence of words) to capture the grammatical properties and richer semantics of the text. For that matter, it combines the scores of different n-grams using the following equation: $CINDER_n\,(C_i, S_i\,) = \frac{1}{m}\sum_{n=1}^{N} w_n\, CINDER_n\,(C_i, S_i\,)$ [PAR2017]

Artificial Intelligence & Information Analysis Lab

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]

Artificial Intelligence & Information Analysis Lab

Video Captioning: Evaluation Metrics - METEOR

METEOR score is calculated as:

Initially, unigram based precision score P is calculated using $P = \frac{m_{cr}}{m_{ct}}$ relationship.

m_{cr}: The number of unigrams co-occurring in both candidate, as well as reference sentences

m_{ct}: The total number of unigrams in the candidate sentences. Then unigram based recall score R is calculated using $R = \frac{m_{cr}}{m_{rt}}$.

[PAR2017]

Video Captioning: Evaluation Metrics - METEOR

m_{rt}: The number of unigrams co-occurring in both candidate as well as reference sentences.

However, m_{rt} is the number of unigrams in the reference sentences. Further, precision and recall scores are used to compute the F-score using following equation: $F_{mean} = \frac{10PR}{R+9P}$ [PAR2017]

Video Captioning: Evaluation Metrics - METEOR

The precision, recall and F-score measures account for unigram based congruity and do not cater for n–grams. The n–gram based similarities are used to calculate the penalty p for alignment between candidate and reference sentences. This penalty takes into account the nonadjacent mappings between the two sentences. [PAR2017]

Video Captioning: Evaluation Metrics - METEOR

The penalty is calculated by grouping the unigrams into minimum number of chunks. The chunk includes unigrams that are adjacent in candidate as well as reference sentences. If a generated sentence is an exact match to the reference sentence then there will be only one chunk. The penalty is computed as

$$p = \frac{1}{2}\left(\frac{N_c}{N_u}\right)^2, \text{[PAR2017]}$$

Video Captioning: Evaluation Metrics - METEOR

N_c : The number of chunks and N_u corresponds to the number of unigrams grouped together. The METEOR score for the sentence is then computed as:

$$M = F_{mean}\,(1 - p)$$

Corpus level score can be computed using the same equation by using aggregated values of all the arguments i.e. P, R and p. In case of multiple reference sentences, the maximum METEOR score of a generated and reference sentence is taken. [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: Evaluation Metrics - F-Score

F-score, also called the **F1-score.** Common adjusted F-scores are the F0.5-score and the F2-score, as well as the standard F1-score. The formula for the standard F1-score is the harmonic mean of the precision and recall. A perfect model has an F-score of 1. A perfect model has an F-score of 1. The Mathematical definition of the F-score is: [HTT2019]

Video Captioning: Evaluation Metrics - F-Score

$$F1 = \frac{2}{\frac{1}{recall} X \frac{1}{precition}} = 2\frac{precision\ \ X\, recall}{precision + \ recall}$$

$$precision = \frac{t_p}{t_p + f_p},\ \ \text{recall} = \frac{t_p}{t_p + f_n}\ \ \text{so}$$

$$F1 = \frac{t_p}{t_p + \frac{1}{2}(f_p + f_n)}$$

Artificial Intelligence & Information Analysis Lab

Video Captioning: Evaluation Metrics - F-Score

where **precision** is the fraction of true positive examples among the examples that the model classified as positive (in other words, the number of true positives divided by the number of false positives plus true positives), **recall**, also known as sensitivity, is the fraction of examples classified as positive, among the total number of positive examples[HTT2019]

Video Captioning: Evaluation Metrics - F-Score

(in other words, the number of true positives divided by the number of true positives plus false negatives)

t_p: the number of true positives classified by the model,

f_p: the number of false positives classified by the model, and

f_n: the number of false negatives classified by the model. [HTT2019]

Video Captioning: Future Directions

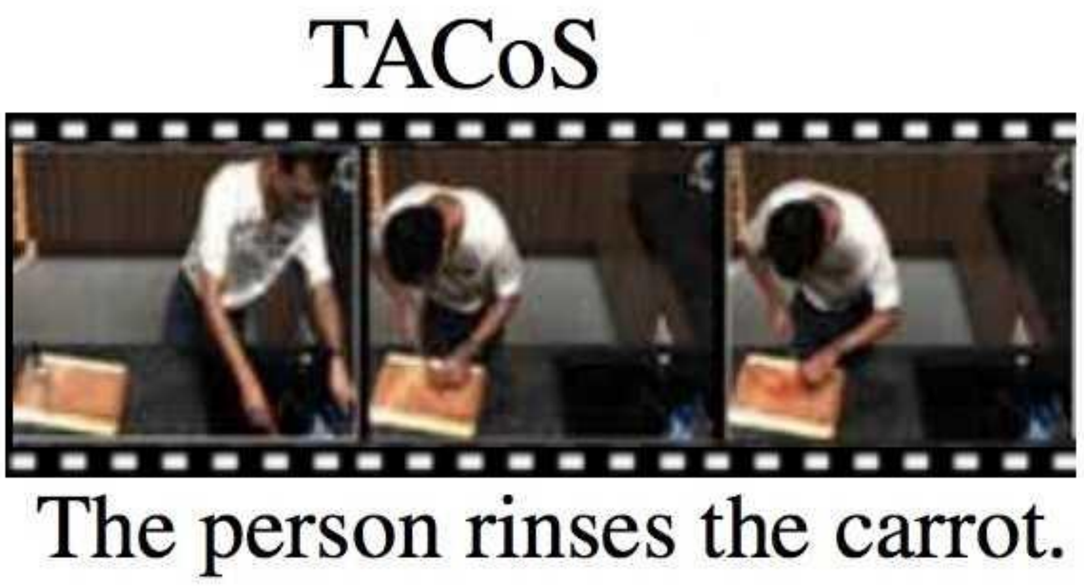
Video captioning problem is not yet solved, as the best performance so far is still far from human-level captioning.

Here, are listed several possible future directions, according to discussions in the literature and progress in related fields:

Video Captioning: Datasets

TACoS Dataset, contains videos of different activities in the cooking domain in an indoor environment. Each video is annotated with both fine-grained activity labels with temporal locations and descriptions with temporal locations by multiple Amazon Mechanical Turkers. It has a total of 18,227 video-sentence pairs on 7,206 unique time intervals. [WET2013]

Video Captioning: Datasets



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

Video Captioning: Datasets

TACoS-Multi dataset is an extension to the dataset with paragraph description per temporal segment, but the limitation is still the same that the setting is closed-domain and too simple for learning. [WET2013]

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

Microsoft Video Description Corpus (MSVD), also referred as YouTube Dataset in early works. It is a collection of YouTube clips collected on Mechanical Turk by requesting workers to pick short clips depicting a single activity. Each clip lasts between 10 seconds to 25 seconds. It has 1,970 videos clips in total and covers a wide range of topics such as sports, animals and music. [CHE2011]

Video Captioning: Datasets

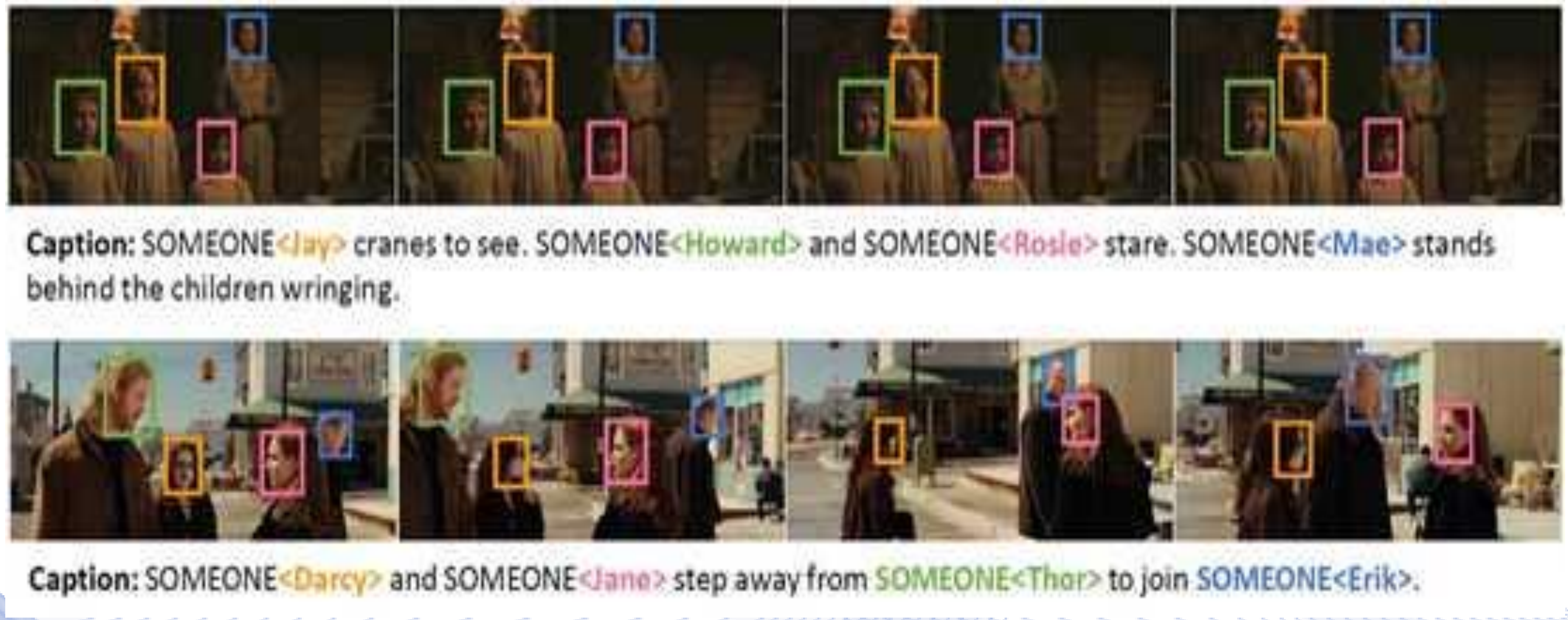
Sentences:
- A man lights a match book on fire.
- A man playing with fire sticks.
- A man lights matches and yells.

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

Video Captioning: Datasets

Montreal Video Annotation Dataset (M-VAD) is a large-scale movie description dataset from the DVD descriptive video service (DVS) narrations. DVS are audio tracks describing the visual elements of a movie, produced to help visually impaired people. The dataset has 49k video clips extracted from 92 DVD movies. [LAR2016]

Video Captioning: Datasets



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

Video Captioning: Datasets

MPII Movie Description Corpus (MPII-MD). It contains around 37,000 movie clips from 55 audio descriptions (ADs) available movies and about 31,000 movie clips of 49 Hollywood movies. Each video clip is equipped with one sentence from movie scripts and one sentence from DVD descriptive video service (DVS). [TAN2015]

Video Captioning: Datasets

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

MSRVideo-to-Text (MSR-VTT). It is by far the largest video captioning dataset in terms of the number of sentences and the size of the vocabulary. It contains 10k video clips crawled from a video search engine from 20 most representative categories of video search, including news, sports etc. The duration of each clip is between 10 and 30 seconds, while the total duration is 41.2 hours. [YAO2016]

Video Captioning: Datasets

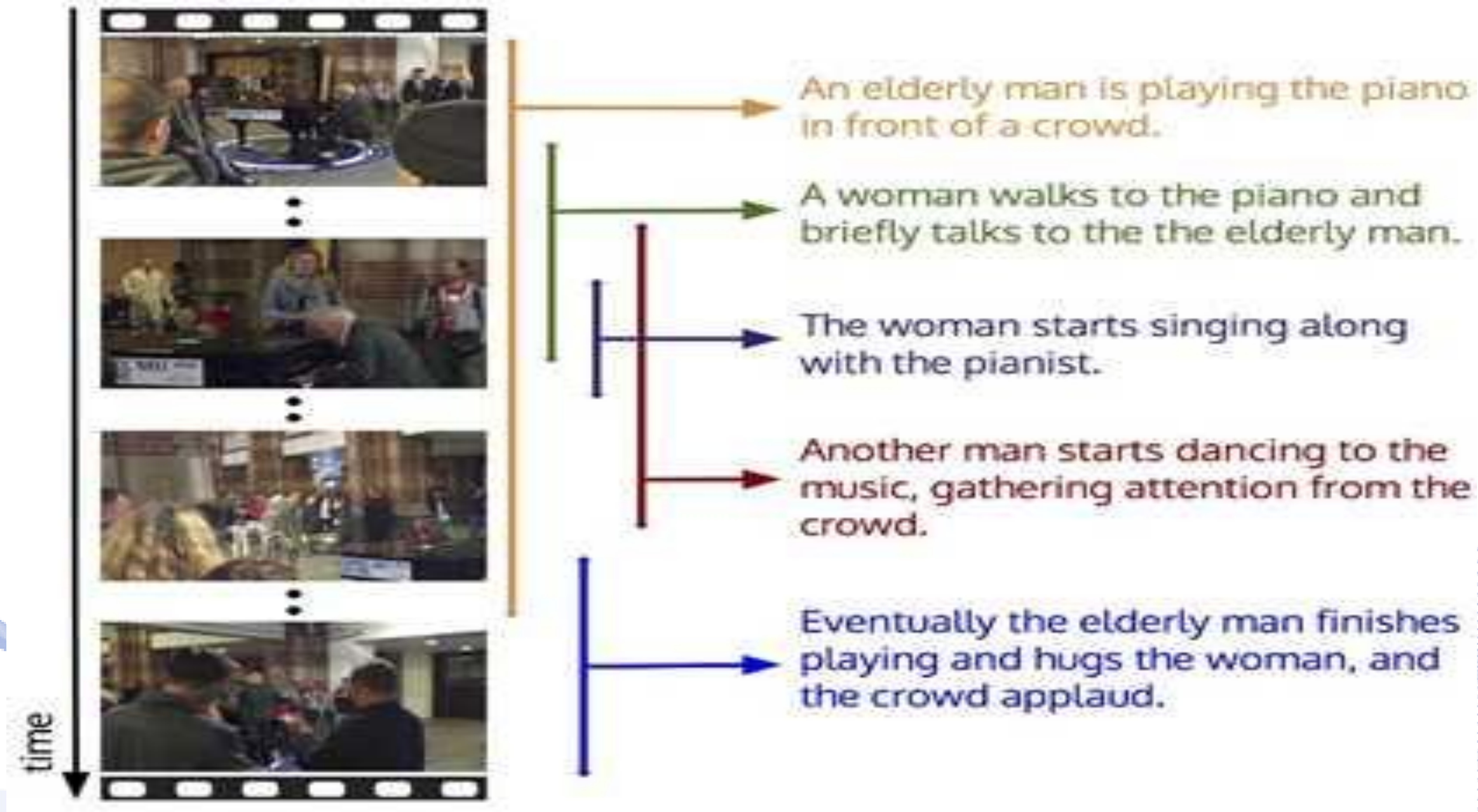
MSR Video-to-Text (MSR-VTT) dataset. (Image from https://paperswithcode.com/dataset/msr-vtt)

Video Captioning: Datasets

ActivityNet Captions is a recently released largescale benchmark dataset specific for dense-captioning events. It contains 20k videos amounting to 849 video hours. The videos are collected from video search engine, covering a wide range of categories. On average, each video contains 3.65 temporally localized sentences, resulting in a total of 100k sentences. [REN2017]

Artificial Intelligence & Information Analysis Lab

Video Captioning: Datasets



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

Artificial Intelligence & Information Analysis Lab

Video Captioning: Datasets

SumMe dataset of 25 personal videos obtained from the YouTube covering holidays, events and sports. They are raw or minimally edited user videos, i.e., they have a high compressibility compared to already edited videos. The length of the videos ranges from about 1 to 6 minutes. The videos are unedited or minimally edited. The dataset provides 15–18 reference summaries for each video. [HAY2014], [YUT2019]

Video Captioning: Datasets

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

Video Captioning: Datasets

TVSum (Title-based Video Summarization), is an unsupervised video summarization framework that uses the video title to find visually important shots.

TVSum contains 50 YouTube videos, each of which has a title and a category label as metadata and their shot level importance scores annotated via crowdsourcing. [YUT2019], [STE2015]

Video Captioning: Datasets

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

Video Captioning: Datasets

Hollywood [SCH2008] and **Hollywood2** [LAPT2005] , are more recent data sets, that attempt to provide a more challenging problem and consist of actions “in the wild” consisting of video clips taken from a variety of Hollywood feature films . These datasets presented a new level of complexity to the recognition community, arising from the natural within-class variation of unconstrained data.

Video Captioning: Datasets

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

Video Captioning: Datasets

Hollywood3D and **Hollywood3D2** are a new natural action data set. They are build on the spirit of the existing Hollywood data sets but includes 3D information. This 3D information gives additional visual cue's which can be used to help simplify the within-class variation of actions. Lighting variations are generally not expressed in depth data, and actor appearance differences are eliminated. [HAD2013]

Video Captioning: Datasets

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

Video Captioning: Future Directions

- Dense captioning
- Attention mechanism
- Audio that accompanies visual frames
- Some works on learning with web image search
- Discovering objects, actions and their interactions
- The temporal structure of video is intrinsically layered

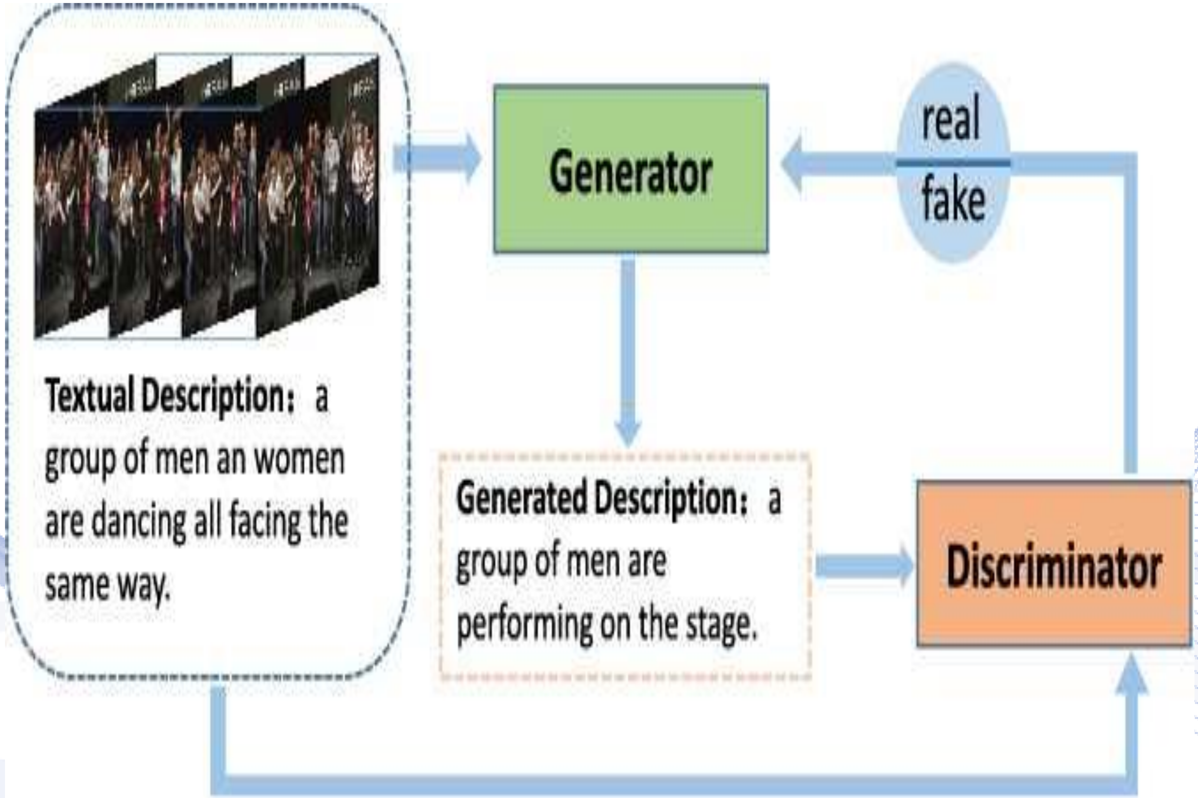
Video Captioning by Adversarial LSTM

Specifically, in this model it has been adopted a standard generative adversarial network (GAN) architecture, characterized by an interplay of two competing processes: a “generator” that generates textual sentences given the visual content of a video and a “discriminator” that controls the accuracy of the generated sentences.

Video Captioning by Adversarial LSTM

In an Adversarial LST model, the discriminator acts as an "adversary" toward the generator, and with its controlling mechanism, it helps the generator to become more accurate. For the generator module, we take an existing video captioning concept using LSTM network.

Video Captioning by Adversarial LSTM



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

For the discriminator, it has been propose a novel realization specifically tuned for the video captioning problem and taking both the sentences and video features as input. This expansion of the LSTM concept enabled the video captioning process to improve the accuracy and diversity of generated captions and their robustness to increasing video length. [ZHO2018]

Video Captioning by Adversarial LSTM

Applying a GAN to the context of video captioning is, however, not straightforward. A GAN is designed for real-valued, but continuous data and may have difficulty handling sequences of discrete words or tokens. The reason lies in that the gradient of the loss from the discriminator based on the output of the generator is used to move the generator to slightly change the way the sentences are generated. [ZHO2018]

Video Captioning by Adversarial LSTM

However, if the output of the generator consists of discrete tokens, the slight change guidance by the discriminator may not work because there may be no token in the used dictionary to signal the desired level of change towards the generator. In order to overcome this problem, it has been proposed an embedding layer which can transform the discrete outputs into a consecutive representation. [ZHO2018]

Video Captioning by Adversarial LSTM

Besides that, since the outputs of the generative model are a sequence, ordinary discriminative model, consisted of several fully connected layers, has a poor ability for classifying the sequence-sentence. For solving this problem, it has been proposed a new realization of the discriminative model. Specifically, it has been replaced the fully connected layer, with a novel convolutional structure. [ZHO2018]

Video Captioning by Adversarial LSTM

This discriminative model consists of convolutional layer, max-pooling layer and fully connected layer. The convolutional layer will produce local features and retain the local coherence around each word of the sequence-sentence. After max-pooling layer, the most important information of the sentence will be effectively extracted. [ZHO2018]

Video Captioning by Adversarial LSTM

Those information are denoted by a fixed length of vector. Additionally, it has been also introduced a multimodal input for the discriminative model. It has been sent not only the sentence to the discriminative module but also the video feature generated from the first LSTM layer (Encoder) of generative module. The novel methods for incorporating the original inputs with the video feature helped to generate more relevant descriptions about the input video. [ZHO2018]

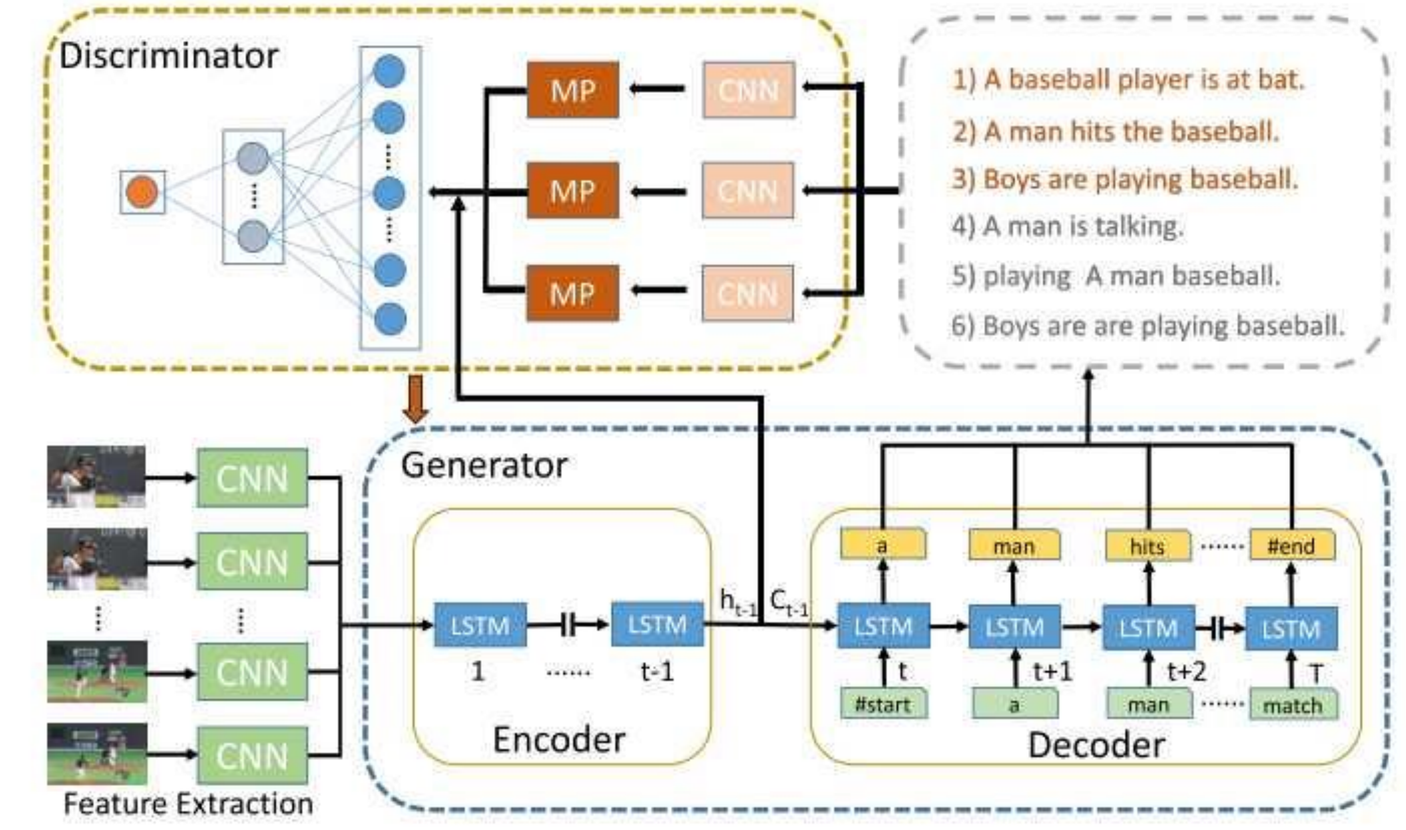
Video Captioning by Adversarial LSTM

Although the LSTM scheme has proved promising performance for handling the temporal nature of video data in the temporal process, the LSTM scheme critical deficiency is shown to accumulate the grammatical errors exponentially and may result in decreasing association among the generated words with the increasing video length. Based on the problem, it is considered if there is a structure that can discriminate whether the generated descriptions are reasonable and relevant to the video. [ZHO2018]

Video Captioning by Adversarial LSTM

Inspired by the generative adversarial network firstly for generating an image, it has been proposed the model LSTM-GAN incorporating a joint LSTMs with adversarial learning. This model consists of a generative model and discriminative model. The generative model is used for encoding the video clips and generates sentences, while the discriminative model is trying to distinguish whether the input sentences are from reference sentence or generated sentences. [ZHO2018]

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]

Video Captioning by Adversarial LSTM example

If the **problem** is:

Given a video V that includes a sequence of n sample frames where $V = \{v_1, v_2, \dots, v_n\}$ with associated caption S where $S = \{w_1, w_2, \dots, w_m\}$ consisting of m words. Let $v_i \in R^{D_v}$ and $w_j \in R^{D_w}$ denote the D-dimensional visual presentations of the i-th frame in video V and D_w-dimensional textual features of the j-th word in sentence S, respectively. [ZHO2018]

Video Captioning by Adversarial LSTM example

In our work, our goal is to maximize the conditional probability of an output sequence $S = \{w_1, w_2, \dots, w_m\}$ given an input sequence $S = \{v_1, \dots, v_m\}$. The conditional probabilities over the sentences can be defined as follows:

$$p(s|v) = p(w_1, \dots, w_m | v_1, \dots, v_n) \text{ [ZHO2018]}$$

Artificial Intelligence & Information Analysis Lab

Video Captioning by Adversarial LSTM example

This problem is similar to the problem of machine translation in natural language processing, where a sequence of words serves as input into a generative model that outputs a sequence of words as the translation result. What is different from aforementioned is that, it has been replace the textual input by the video frames and look forward to a sequence of caption as output. What is more, it is not only expected to get the relevant description of the input videos but also to make the sentences natural and reasonable for people to understand. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **Proposed Solution** is:

The model consists of a generative model G and discriminative model D. The generative model G, defines the policy that generates a sequence of the relevant description given a short video.

The discriminative model D is a binary classifier that takes a sequence of sentences $\{s, y\}$ as input and outputs a label $D(S) \in [0, 1]$ indicating whether the sentence is natural, reasonable and grammatical correct. [ZHO2018]

Artificial Intelligence & Information Analysis Lab

Video Captioning by Adversarial LSTM example

The **designed architecture** is:[ZHO2018]

1. **Objective Function**:

In order to achieve faster convergence of the objective, we firstly pre-training the generative model G and the discriminative model D, respectively. For G, similar to sequence to-sequence models, our goal is to estimate the conditional probability $p(S|V)$ where $V = \{v_1, v_2, \ldots, v_t\}$ is an input sequence consisting of a sample of frames and $S = \{w_1, w_2, \ldots, w_t\}$ is the corresponding output sequence as a descriptive texture for the input video.

Video Captioning by Adversarial LSTM example

The **designed architecture** is:[ZHO2018]

1. **Objective Function**:

t: The length of the video

t_1: the input sentence.

As sequence-to-sequence models, we conclude the follow objective function:

$$p(s|v) = p(w_1, \dots, w_m | v_1, \dots, v_n) = \Pi_{t=1}^{t_1} p(w_i | V, w_1, \dots, w_i - 1).$$

Video Captioning by Adversarial LSTM example

The **designed architecture** is:[ZHO2018]

1. **Objective Function**:

For D, our primary purpose is to train a classifier which can be used for sentence encoding and mapping the input sentence to an output $D(S)$ € $[0,1]$ representing the probability of S is from the ground truth-captions, rather than from adversarial generator.

The objective function of D for pre-training can be formalized into a cross-entropy loss as follow:

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

1. **Objective Function**:

$$L_D(Y, D(S)) = -\frac{1}{m}\sum_{i=1}^{m}\left[(Y_i)log\big(D(S_i)\big) + (Y_i)log(1 - D(S_i))\right]$$

m: The number of examples in a batch,

$Y_i D(S_i)$: The real label and predicted value of discriminator respectively. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

2. Generative Model:

We use a joint recurrent neural networks, also called encoder-decoder LSTM similar to sequence-to-sequence models, as the generative model. The encoder architecture is used to encode the video features into a fixed dimension vector. While the decoder architecture decodes the vector into natural sentences. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

2. Generative Model:

To begin with, we adopt VGG16 the sequence frames as the CNN architecture to map $V = \{v_1, v_2, \ldots, v_t\}$ into a feature matrix $W_v \in R^{D_d X D_t} = \{wd_1, \ldots, w_{Dt}\}$. D_d and D_t denote the dimensions of a feature vector and the number of frames, respectively. The encoder LSTM net, maps the input embedding presentations namely features matrix, into a sequence of hidden states $h_1, h_2, \ldots, h_t$. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

2. Generative Model:

h_t: The last status, as the presentations of the whole video, generated from “encoder”, will be sent to the decoder LSTM which is referred to as “decoder”.

We adopt a soft-argmax function:

$W_t - 1 = \varepsilon_{we}(softmax(Vh_{t-1} \odot L), W_e)$ [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

2. Generative Model:

$W_e \in R^{ZXC}$: A word embedding matrix (to be learned) and transforms the one-hot encoding of words to a dense lower dimensional embedding,

C: The dimension of the embedded word and

Z: The size of vocabulary in our training data.

V: The set of parameters and encodes the h_{t-1} to a vector. W_{t-1}: The generated word of LSTM at t^{th} step. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:[ZHO2018]

2. Generative Model:

L is a big enough integer which would make the vector of Softmax $(Vh_{t-1} \odot L)$ closes to a one-hot form. Each value of it is constrained to be either approximately 0 or 1 which can help the

W_{t-1} more close to $W_e[t-1]$ (suppose the value $(t-1)$ position is the largest of Vh_{t-1}) and also help the word embedding to be more smooth and speed up the loss function to convergence. ε denotes a function that maps the decoder output space to a word space.

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

3. Discriminative Model:

In the discriminator D, our primary purpose is to maximize the probability of assigning the correct label to both training sentences and generated sentences from G. The discriminator consists of a convolution layer and a max-pooling operation, which can capture the most useful local features produced by the convolutional layers, over the entire sentence for each feature map. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

3. **Discriminative Model:**

The input sentences to our discriminator contain both the ground-truth sentences as the true label and generated sentences generated by our generator as the false label. For convenience, we fix the length of input sentences by adopting the length of longest sentence in a mini-batch (padded 0 when necessary). [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

3. Discriminative Model:

A sentence of length T is represented as a matrix $X_d \in R^{CXT} = \{x_{d_1}, \ldots, x_{d_T}\}$ by concatenating the word embeddings as columns, where T is the length of sentence and C is the dimension of a word. Then a kernel $W_c \in R^{CXl}$ applies a convolution operation to a window size of T words to produce a feature map as one of the representations of the input sentence. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

3. Discriminative Model:

This process could be formulated as follow:

$$Out = f(X * W_c + b) \in R^{T-l+1}$$

$f(X * W_c)$: A nonlinear activation function (for example RELU),

$b \in R^{T-l+1}$: The bias vector and $*$ represents the convolution operator. [ZHO2018]

Video Captioning by Adversarial LSTM example

The **designed architecture** is:

3. Discriminative Model:

To verify the impressive performance of our video captioning by adversarial training approach, we can evaluate and compare our experimental results on four large public datasets, including MSVD, MSR-VTT, M-VAD and MPII-MD. [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. e task is naturally decomposed into two sub-tasks.

Deep Learning for Video Captioning

One is to encode a video via a thorough understanding and learn visual representation. The other is caption generation, which decodes the learned representation into a sequential sentence, word by word. [YAO2019]

Deep Learning for Video Captioning

Visual perception and language expression are two key capabilities of human intelligence, and video captioning is a perfect example towards learning from human to bridge vision and language.

The goal of video captioning is to automatically describe the visual content of a video with natural language. [YAO2019]

Deep Learning for Video Captioning

Practical applications of automatic caption generation include leveraging descriptions for video indexing or retrieval, and helping those with visual impairments by transforming visual signals into information that can be communicated via text-to-speech technology. Video captioning has already received intensive research attention before the prevalence of deep learning. [YAO2019]

Deep Learning for Video Captioning

At the early stage, video captioning approaches first detect visual concepts in a video with hand-crafted features and then generate the sentence based on pre-defined templates. Such methods highly depend on the templates and the generated sentences are always with fixed syntactical structures, not to mention that the design of hand-crafted features is also bounded for video understanding. [YAO2019]

Deep Learning for Video Captioning

Instead, current deep learning based video captioning often performs sequence to sequence learning in an encoder-decoder paradigm. In between, an encoder equipped with powerful deep neural networks is exploited to learn video representation. A decoder of sentence generation is utilized to translate the learned representation into a sentence with more flexible structures. [YAO2019]

Artificial Intelligence & Information Analysis Lab

Deep Learning for Video Captioning

The learning of video representation is the basis of video understanding, and in general involves both feature extraction and aggregation. The ultimate goal is to extract features from multiple modalities, and then aggregate them spatially and temporally to produce a compact representation. [YAO2019]

Deep Learning for Video Captioning

The recent advances in 2D and 3D Convolutional Neural Networks (CNNs) have successfully improved the state-of-the-art of representation learning from visual, audio and motion information. Nevertheless, feature aggregation particularly for video captioning remains an open challenge. Several techniques from different perspectives, e.g., spatially, temporally and modality-wise, have been studied for exploring feature aggregation in video captioning. [YAO2019]

Deep Learning for Video Captioning

The decoder of sentence generation shares the same learning objectives and evaluation metrics with the sequence generation tasks in NLP field such as text summarization and machine translation. As such, challenges, e.g., exposure bias and objective mismatch, also exist for the decoder in video captioning due to the recursive nature. [YAO2019]

Deep Learning for Video Captioning

Though there are some methods proposed in NLP area , to solve the issues, the complexity of video content and relatively small captioning corpus make it difficult if directly applying these solutions to video captioning. Furthermore, considering that videos in real life are usually long, how to recapitulate all the video content that are worthy of mention is still a valid question. [YAO2019]

Artificial Intelligence & Information Analysis Lab

Deep Learning for Video Captioning example

- **Problem Formulation:**

Given an input video $V = \{f_1, \ldots, f_N\}$ (N: the length of frame sequence), the target of video captioning is to generate a sentence (i.e., word sequence) Y= $\{y_1, \ldots, y_T\}$ to describe the video's content. Thus, video captioning task is often tackled as a problem of sequence-to-sequence learning. [YAO2019]

Deep Learning for Video Captioning example

- **Problem Formulation:**

Most video captioning frameworks are designed as an *encoder-decoder* structure, where the encoder learns condensed video representation from multi-modal features and the decoder produces sentence word-by-word depending on the learned representation from encoder. [YAO2019]

Deep Learning for Video Captioning example

- **Problem Formulation:**

To model the video content, we firstly extract features from multiple modalities:

F= $\{F_V, F_M, F_A, F_S\}$ where $F_V, F_M, F_A, and\ F_S$ denote visual, motion, audio and semantic features respectively. [YAO2019]

Deep Learning for Video Captioning example

- **Problem Formulation:**

To model the video content, we firstly extract features from multiple modalities:

$$F = f_{feat}(V)$$

where $f_{feat}(V)$ is an ensemble of feature extraction functions (usually pre-trained deep neural networks) for multiple modalities of the video. [YAO2019]

Deep Learning for Video Captioning example

- **Problem Formulation:**

The features F may be further aggregated into a more condensed representation, and the process of feature aggregation is conducted depending on some changing state:

$$F_t = f_{aggr}(F, s_t)$$

Where f_{aggr} is the feature aggregation function, s_t is an optional state vector [YAO2019]

Deep Learning for Video Captioning example

- **Problem Formulation:**

(e.g. the model's state when generating the t-th word) and F_t is the aggregated feature. f_{feat} and f_{aggr} constitute the encoder. The *language model* (or decoder) then takes F_t (and optionally and $F_t = s_t$) and predicts the distribution of the word y_t:[YAO2019]

Deep Learning for Video Captioning example

- **Problem Formulation:**

$$p_t = f_{lang}(F_t, s_t)$$

f_{lang}: The updating function in LSTM [Hochreiter and Schmidhuber, 1997] or its variants.

The final prediction of Y is obtained based on the distributions $\{p_1, \dots, p_T\}$. [YAO2019]

Deep Learning for Video Captioning example

- **Video Representation:**

The process to obtain video representation can be divided into two major steps:

Feature Extraction and Feature Aggregation. These methods are also applicable to other video understanding tasks. [YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**

A good set of features is the foundation of a performant video captioning method. Deep learning has been successfully applied to multiple modalities where sufficient amount of data is available, and the learned representations have nice transferability so that they can be directly leveraged by other tasks. [YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Visual**

Visual appearance is the most important feature for understanding video contents. State of-the-art convolutional neural networks (CNNs) have surpassed human performance in recognizing images. [YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Visual**

Activation vectors from higher layers of a trained CNN can capture global visual appearance of its input image, and is now used as the default feature for video captioning. Popular choices of CNN are VGG Net, ResNet and Inception Networks[YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Visual**

Motion feature is crucial for capturing the action sand temporal interactions in video, which complements the static visual appearance. [YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Visual**

3D CNN such as C3D learns spatiotemporal feature by processing a consecutive sequence of video frames with 3-dimensional convolutions, and can selectively attend to both motion and appearance. Thus, the higher-layer activation vectors of 3D CNNare commonly leveraged as motion feature for video captioning[YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Audio**

Audio feature is helpful for distinguishing events such as “person talking to the phone” and “person listening to the phone playing music”. MFCCs (Mel Frequency Cepstral Coefficients) is a widely adopted audio feature, and video captioning works. [YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Semantic**

Semantic feature refers to a wide category of features that explicitly capture semantic contents in videos. MMVD shows that the video-level category information can boost video captioning. Simply incorporating category information into the encoder can yield better captioning performance. [YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Semantic**

MMTGM further predicts latent topics from multimodal features (except semantic feature), then integrates the predicted topics into the designed topic-aware decoder. LSTM-TSA adopts the weakly-supervised attribute detection method to detect frame- and video level fine-grained attributes. [YAO2019]

Deep Learning for Video Captioning example

- **Multimodal Feature Extraction:**
  - **Semantic**

Next a transfer unit is utilized to dynamically incorporate attribute information into LSTMbased decoder. In this sense, semantic features of any granularity can improve video captioning, which is because they provide the decoder (language model) with more prior knowledge about the video content. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Temporal Attention**

Next a transfer unit is utilized to dynamically incorporate attribute information into LSTMbased decoder. In this sense, semantic features of any granularity can improve video captioning, which is because they provide the decoder (language model) with more prior knowledge about the video content. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Temporal Attention**

The simplest way to aggregate a feature sequence, is using a LSTM/GRU to encode the sequence and take the final encoding state as the aggregated feature for decoding. However, treating video features as a that sequence is not effective, because:[YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Temporal Attention**

(1) the length of gradient flow to the earliest frame is as long as the sequence, which leads to gradient vanishing; **(2)** each feature in the sequence contributes the same to the decoder, which makes the model also pay attention to background noises.

Temporal attention (also known as dynamic attention) [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Temporal Attention**

Temporal attention (also known as dynamic attention) is a mechanism which learns to dynamically assign weights to each feature in the sequence such that the decoder can pay more attention to relevant features when generating certain words. Thus the computation of attention weights involves both visual feature sequence and the decoder state. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Temporal Attention**

Temporal attention (also known as dynamic attention) is a mechanism which learns to dynamically assign weights to each feature in the sequence such that the decoder can pay more attention to relevant features when generating certain words. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Temporal Attention**

Thus the computation of attention weights involves both visual feature sequence and the decoder state. Another effect is that the decoder and each feature is directly connected by a weighted path, which shortens the length of gradient flow and leads to more effective learning. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Temporal Attention**

hLSTMat is an improved temporal attention mechanism which makes the decoder depend less on visual features when generating non-visual words, but instead rely on language model's state. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Spatial Attention**

Different regions of the video frames also contribute differently to the final word prediction, e.g. objects are clearly more important than background. Spatial attention methods aim to learn spatial attention maps, which indicate the importance of different regions. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
    - **Spatial Attention**

Dynamic attention can also be applied spatially if regions are treated sequentially. Thus, MAM-RNN adopts two-level spatial and temporal dynamic attention for video captioning. When computing spatial attention weights for a certain frame, MAMRNN additionally incorporates the attention weights from previous frame. In this way, the spatial attention maps are linked across time. [YAO2019]

Deep Learning for Video Captioning example

- **Feature Aggregation:**
  - **Multimodal Feature Fusion**

Using multimodal features is ubiquitous in video captioning methods, in contrast, multimodal feature fusion strategy is rarely explored. MMVD simply concatenates features from multiple modalities as the input to decoder. It is obvious that the importance of each modality is different for various types of videos. [YAO2019]

Deep Learning for Video Captioning example

Caption Generation

Given the generated word probabilities at each time step $\{p_1, \dots, p_T\}$ and ground truth caption $\hat{Y} = \{\hat{y}_1\ , \dots, \hat{y}_T\}$ the most common learning objective for captioning is to maximize the loglikelihood

of all the ground truth words: $max_\theta = \sum_t^T = p_t(\hat{y}_T)$, where θ is all the learnable parameters of the captioning model. [YAO2019]

Deep Learning for Video Captioning example

Caption Generation

This objective is widely adopted for sequence generation tasks such as machine translation and captioning. However, there are two major problems with it. First, there is a discrepancy between this objective function and the automatic evaluation metrics such as BLEU. [YAO2019]

Deep Learning for Video Captioning example

Caption Generation

This is often referred to as objective mismatch. And there is also a gap between these metrics and human judgment. Second, this objective alone maybe insufficient to train a good language model since video captioning datasets have a much smaller corpus compared to pure NLP datasets. [YAO2019]

Bibliography

[YANG2018] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, H.T. Shen, Y. Ji Y, “Video captioning by adversarial LSTM”, IEEE Transactions on Image Processing, 27(11):5600-11, 2018.

[HTT2020] “What is the use of captions”. https://askinglot.com/what-is-the-use-of-captions, 2020

[HTT2013] “Howto captions cutlines”. https://web.ku.edu/ edit/captions.html, 2013

[LED2018] JACLYN LEDUC. “3 reasons why captioning is more important now than ever before”. https://www.3playmedia.com/blog/importance-of-captioning/, 2018.

[HTT2019] “How video captions help attract and engage more users”

https://www.lemonlight.com/blog/how-video-captions-help-attract-and-engage-moreusers/, 2019

[THO2020] Thoudam Doren, Singh Alok Singh, Sivaji Bandyopadhyay. “NITS-VC system for VATEX Video Captioning”, Challenge 2020. Center for Natural Language Processing Department of Computer Science and Engineering National Institute of Technology Silchar

Assam, India:arxiv. 2006.04058v2. September 2020.

Artificial Intelligence & Information Analysis Lab

Bibliography

[PRA2013] Pradipto Das, Chenliang Xu, Richard F. Doell, Jason J. Corso, “A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching”, IEEE Conference on Computer Vision and Pattern Recognition, 2013

[ZAC2012] Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell Waggoner, Song Wang, Jinlian Wei, Yifan Yin, Zhiqi Zhang, “Video in sentences out”, preprint, arXiv:1204.2742, 2012.

[YAN2016] Yi Yang, Fei Wu, Pingbo Pan, Zhongwen Xu, Yueting Zhuang. “Hierarchical recurrent neural encoder for video representation with application to captioning”. pages 1029–1038, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[JIA2018] Jiaqi Su. “Study of Video Captioning Problem”. University, Princeton, 2018.

[HUA2021] Huanhou Xiao, Jinglun Shi. “Diverse video captioning through latent variable expansion”. [cs.CV], 15, June 2021.

Bibliography

[HTT2021] Described and Captioned Media Program. “Captioning types, methods, and styles”.

https://dcmp.org/learn/38-captioning-types-methods-and-styles, 2021.

[WET2013] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, Manfred Pinkal, “Grounding action descriptions in videos”. Transactions of the Association for Computational Linguistics, Volume 1 (TACL), 1:25–36, 2013.

[CHE2011] D. L. Chen, W. B. Dolan. “Collecting highly parallel data for paraphrase evaluation”, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 41, pages:190–200, 2011.

[LAR2016] H. Larochelle, A. Torabi, C. Pal, A. Courville. “Using descriptive video services to create a large data source for video annotation research”, preprint, 2015, arXiv,1503.01070.

[TAN2015] N. Tandon, A. Rohrbach, M. Rohrbach, B. Schiele.” A dataset for movie description”, pages 3202–3212, 2015.

[YAO2016] T. Yao. J. Xu, T. Mei, Y. Rui. “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language”, pages 5288–5296, 2016.

Bibliography

[REN2017] F. Ren, L. Fei-Fei, R. Krishna, K. Hata and J. C. Niebles. “Dense-captioning events in videos”, volume 1, page 6, arXiv:1705.00754v1 [cs.CV] 2 May 2017.

[HAY2014] Hayko Riemenschneider, Michael Gygli, Helmut Grabner, Luc Van. “Creating summaries from user videos”, European Conference on Computer Vision, 2014.

[YUT2019] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, “Rethinking the Evaluation of Video Summaries”, 11 Apr 2019, arXiv:1903.11328v2 [cs.CV].

[STE2015] Amanda Stent, Yale Song, Jordi Vallmitjana, Alejandro Jaimes, “TVSum: Summarizing Web Videos Using Titles”, Yahoo Labs, New York, 2015

[SCH2008] C. Schmid, I. Laptev, M. Marszalek, B. Rozenfeld. “Learning realistic human actions from movies”, In CVPR, 2008.

[LAPT2005] I. Laptev, M. Marszalek,C. Schmid. “Actions in context”, 2005.

[HAD2013] Simon Hadfield, Richard Bowden. “Hollywood 3d: Recognizing actions in 3d natural

Scenes”, Centre for Vision, Speech and Signal Processing University of Surrey, Guildford, Surrey, UK, GU2 7XH, 2013

Bibliography

[DON2014] J. Donahue, M. Rohrbach-R, Mooney S. Venugopalan, H. Xu and K. Saenko. “Translating videos to natural language using deep recurrent neural networks”, preprint, 2014, arXiv 1412.4729

[LIU2019] Wei Liu, Nayyer Aafaq, Ajmal Mian, Syed Zulqarnain Gilani, Mubarak Shah. “Video description: A survey of methods, datasets, and evaluation metrics”, 52(6):1–37, 2019.

[PAR2017] C. Song, J. Park, J. h Han. “A study of evaluation metrics and datasets for video captioning”, In Intelligent Informatics and Biomedical Sciences (ICIIBMS), International Conference on pages. IEEE, pages 172–175, 2017.

[HTT2019] Thomas Wood. “F-score”, https://deepai.org/machine-learning-glossary-and-terms/f-score, 2019.

[MOO2015] J. Donahue R. Moone T. Darrell S. Venugopalan, M. Rohrbach, K. Saenko, “Sequence to sequence-video to text”, pages 4534–4542, In Proceedings of the IEEE international conference on computer vision, 2015.

Bibliography

[SHI2016] K. Ohnishi, A. Shin and T. Harada. “Beyond caption to narrative: Video captioning with multiple sentences”, Image Processing (ICIP), 2016 IEEE International Conference on

IEEE, pages 3364–3368, 2016.

[QIU2014] W. Qiu, A. Friedrich, M. Pinkal, A. Rohrbach, M. Rohrbach and B. Schiele. “Coherent multisentence video description with variable level of detail”, pages 184–195, In German conference on pattern recognition.Springer, 2014.

[HUA2016] Z. Huang, Y. Yang, H. Yu, J. Wang and W. Xu. “Video paragraph captioning using hierarchical recurrent neural networks”, pages 4584–4593, In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[ZHO2018] Yang Y., Zhou J., Ai J., Bin Y., Hanjalic A., & Shen H. T, “Video Captioning by Adversarial LSTM. IEEE Transactions on Image Processing, 27(11), 5600-5611, 2018.

[YAO2019] Ting Yao, Shaoxiang Chen, Yu-Gang Jiang. “Deep learning for video captioning: A Review”, Shanghai Key Lab of Intelligent Info, Processing, School of Computer Science, Fudan University, China., Jilian Technology Group (Video++), Shanghai, China., JD AI Research, China, 2019

Artificial Intelligence & Information Analysis Lab

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