Video Description

Ir. He Ming Zhang
Advisor: Prof. C.-C. Jay Kuo

Outline

- Motivation
- Problem definition
- Preliminaries
- Related works
- Conclusion

Outline

- **Motivation**
- Problem definition
- Preliminaries
- Related works
- Conclusion

Motivation

We have ...

- **huge amount of video**

  Every minute, 100 hours of video are uploaded to YouTube[1].

We lack ...

- **time to watch all the videos**
- **description of videos**

We want ...

- **computer to <u>understand</u> the visual content**
- **computer to <u>describe</u> the visual content**

[1]https://www.youtube.com/yt/press/statistics.html accessed on 2015-02-06.

Motivation

Applications

- **Tagging**

woman | dog

VS

A woman is walking a dog

A woman is chased by a dog

- **Indexing**

  Improving indexing and search quality for online videos.

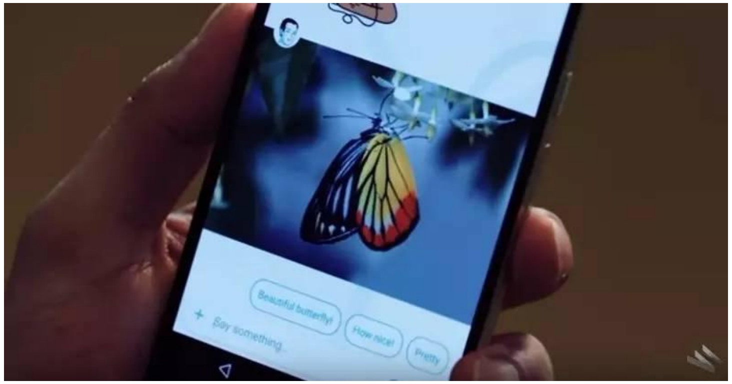
Motivation

Applications

- **Human-robot interaction**

  Describing movies for the blind

  As well as for the lazy people...



Outline

- Motivation
- Problem definition
  - Problem for researchers
  - Datasets
  - Evaluation
- Preliminaries
- Related works
- Conclusion

Problem Definition

Problem for researchers

- **From video clip to natural language**
  - **Input - video clip**

    Typically from several to few tens of seconds
    A specific domain or open domain (“in the wild”)

  - **Output - natural language that describes the content of the input**

    One or more sentence(s) in natural language (usually in English)

- **Different from image description**
  - **Video contains more information**

    more or less difficult?

USC Viterbi
School of Engineering

Problem definition

Datasets

Dataset	multi-sentence	domain	sentence source	vides	clips	sentences
YouCook [1]	x	cooking	crowd	88	-	2668
TACoS [2]	x	cooking	crowd	127	7206	18227
TACoS Multi-Level [3]	x	cooking	crowd	185	14105	52593
MSVD [4]	o	open	crowd	-	1970	70028
MVAD [5]	x	open	professional	92	48986	55904
MPII-MD [6]	x	open	professional	94	68337	68375

Problem Definition

Datasets

- **Trend - more challenging**
  - **Broader domains**
    From single domain to open domain
  - **Larger datasets**
    More sentences/ clips

Problem Definition

Datasets

- **MSVD**
  - **YouTube videos**

    e.g. from 0:33 to 0:46,
    http://www.youtube.com/watch?v=mv89psg6zh4

  - **Multi-descriptions**
    - A bird in a sink keeps getting under the running water from a faucet.
    - A bird is bathing in a sink.
    - A bird is splashing around under a running faucet.
    - A bird is standing in a sink drinking water that is pouring out of the faucet.
    - ...

Problem Definition

Datasets

- **MSVD**
  - **YouTube videos**

    e.g. from 0:11 to 0:14,
    http://www.youtube.com/watch?v=cSDkshD2ME0

  - **Multi-descriptions**
    - Someone behind a rock shoots a man on horseback who slumps forward onto his horse.
    - A man shoots a man on a horse.
    - A man hiding behind a rock shoots a man on horseback with a rifle.
    - A man is shooting another man.
    - ...

Problem Definition

[1] Das, Pradipto, et al. "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[2] Regneri, Michaela, et al. "Grounding action descriptions in videos." Transactions of the Association for Computational Linguistics 1 (2013): 25-36.

[3] Rohrbach, Anna, et al. "Coherent multi-sentence video description with variable level of detail." Pattern Recognition. Springer International Publishing, 2014. 184-195.

[4] Chen, David L., and William B. Dolan. "Collecting highly parallel data for paraphrase evaluation." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[5] Torabi, Atousa, et al. "Using descriptive video services to create a large data source for video annotation research." arXiv preprint arXiv:1503.01070 (2015).

[6] Rohrbach, Anna, et al. "A dataset for movie description." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

Problem Definition

Example results from state-of-the-art [7]

A man is pouring oil into a pot.

A dog is playing in a bowl.

[7] Yu, Haonan, et al. "Video Paragraph Captioning using Hierarchical Recurrent Neural Networks." CVPR 2016.

Problem Definition

Evaluation

- **Difficulties**

  Natural language is rich

  Description may be partially wrong/correct

  No standard metric (a few metrics are used by different researchers)

Problem Definition

Evaluation

- **Methods**

  Human evaluation

  Binary rating (correct/ incorrect)

  Scale rating (e.g. 1~5)

Problem Definition

Evaluation

- **Methods**

  Automated evaluation:

  BLEU (BiLingual Evaluation Understudy)

  – one of the first metrics to achieve a **high correlation with human judgements** of quality
  – **modified version of F-score**
  – example:

  Ref: Israeli officials are responsible for airport security.
  A: Israeli officials responsibility of airport safety.
  B: Airport security Israeli officials are responsible.
  Score: A - 0%
  B - 52%

Problem Definition

Evaluation

- **Methods**

  Automated evaluation:

  METEOR (Metric for Evaluation of Translation with Explicit ORdering)

  - **higher correlation with human judgements** in both corpus and sentence level
  - **modified version of F-score**
  - flexible matching (partial credit)

    Ref: Joe goes home

    A: Jim went home

    B: Jim walks home

Outline

- Motivation
- Problem definition
- Preliminaries
  - Statistical Machine Translation (SMT)
  - Recurrent Neural Network (RNN)
- Related works
- Conclusion

Preliminaries

We need ...

- **recognition (CRF, CNN, etc)**
  - objects
  - scene / background
  - events

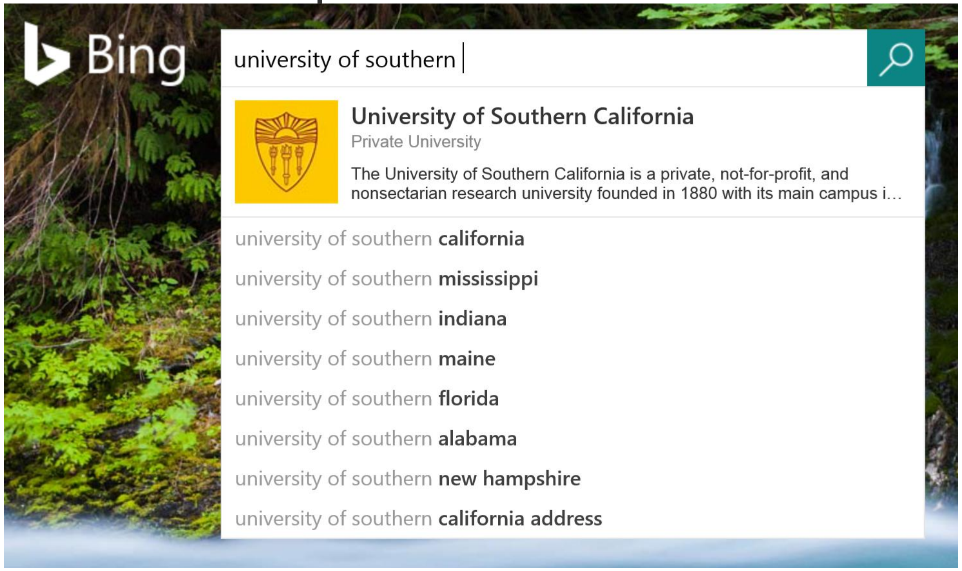
- **language processing (manual rules, SMT, RNN, etc)**
  - word selection
  - sentence generation

Preliminaries

n-gram

- **Markov model with higher order**

  In a language model, the probability of a word is conditioned on **some number** of previous words.



- **Properties and usages**

  It is used in **statistical natural language processing**.

Preliminaries

Statistical Machine Translation (SMT)

- **Statistical model**

  It **translates** the document according to the probability distribution *p(T|S);*

  Examples:

  - Word-level

    S (Dutch): Ik ben een promovendus.

    T (English): I am a PhD student.

  - Semantic-level

    S (Dutch): Ik ben het er mee eens.

    T (English): I am it here with in agreement.

    T (English): I agree with it.

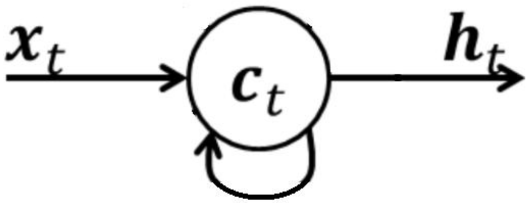
  The system can not store all native strings and their translation, therefore the language models are approximated by n-gram models.

Preliminaries

Recurrent Neural Network (RNN)

- **Internal memory**

  A class of neural network where connections between units form a **directed cycle**;



- **Properties and usages**

  It can **process sequential data** and be used for **language modeling, handwriting recognition**, etc

  Traditional RNNs are **very hard to train**;

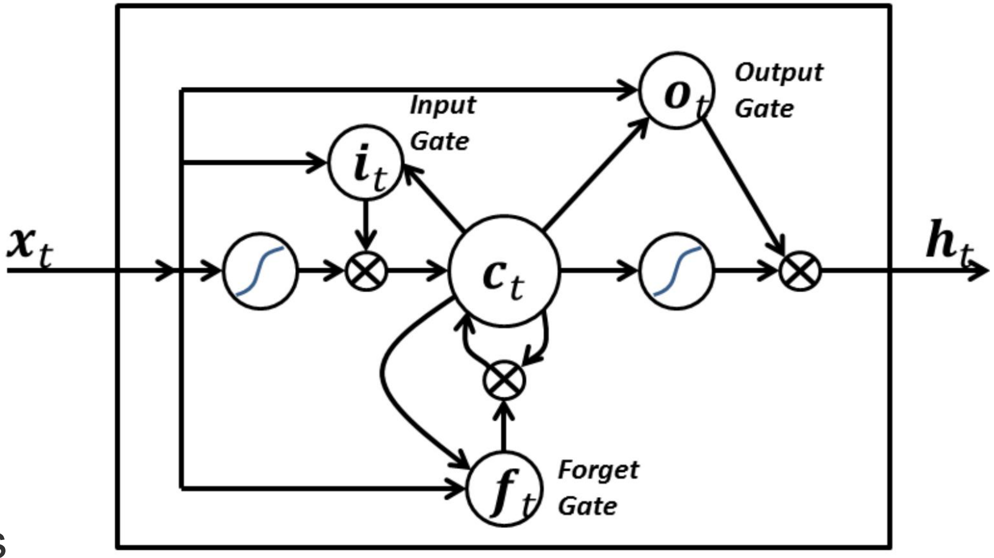
Preliminaries

Recurrent Neural Network (RNN)

- **LSTM (Long Short-Term Memory)**

Internal memory for an arbitrary length of time;

- **Input** gate: determines when the unit should let the input flow into its memory
- **Forget** gate: determines when the unit should forget the value in its memory;
- **Output** gate: determines when the unit should output the value in its memory.



A LSTM unit [8]

[8] Greff, Klaus, et al. "LSTM: A search space odyssey." arXiv preprint arXiv:1503.04069(2015).

Outline

- Motivation
- Problem definition
- Preliminaries
- Related works
  - Early works
  - Recent works
  - Summary
- Conclusion

Related works

Early works

- **Youtube2text [9]**

  Mine *(Subject, Verb, Object)* triplets from the natural language descriptions of the videos

  Build a separate semantic hierarchy for each part of the triplet *(H_S, H_V , and H_O)*.

  Dectect objects and activities using existing object and motion descriptors

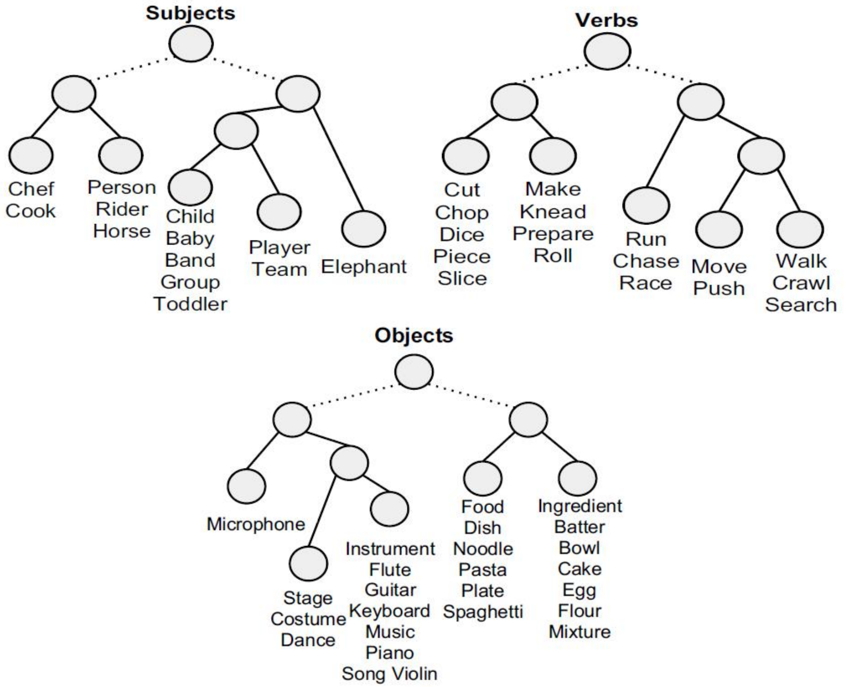


Figure 3: Small portions of the Hierarchies learned over Subjects, Verbs and Objects

[9] Guadarrama, Sergio, et al. "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition." Proceedings of the IEEE International Conference on Computer Vision. 2013.

Related works

Early works

- **Youtube2text**

Language model

- For activities that are unseen during training, they expand detected verbs with similar verbs.
  e.g. for (person, move, car), expand "move" with "ride" and "drive" without training videos for "ride" or "drive"
- Select the best triplet

$$score = p(S \mid video) * p(V_{expand} \mid video) * Similarity(V_{expand}, V_{original}) * p(O \mid video) * SVO\_likelihood$$

- Generate sentences using **manual template**

Related works

Early works

- **Youtube2text**

Experimental results on MSVD

Automated evaluation

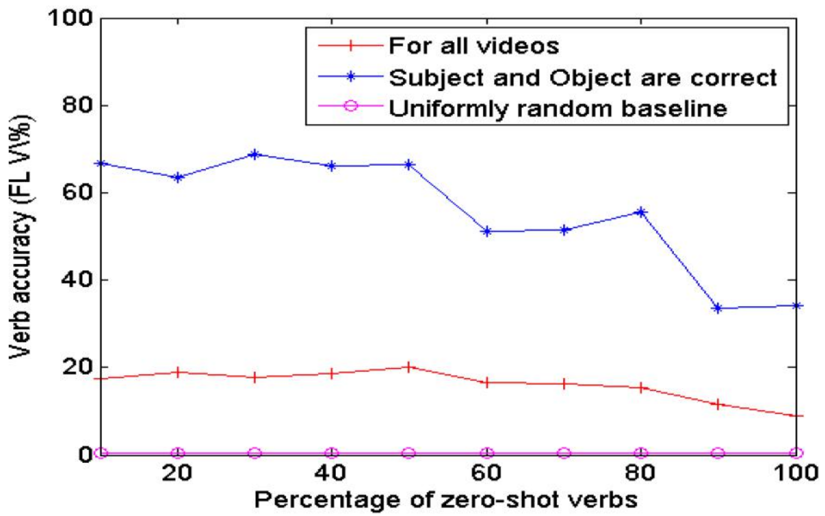


Figure 4: Zero-shot Activity Recognition

Human evaluation

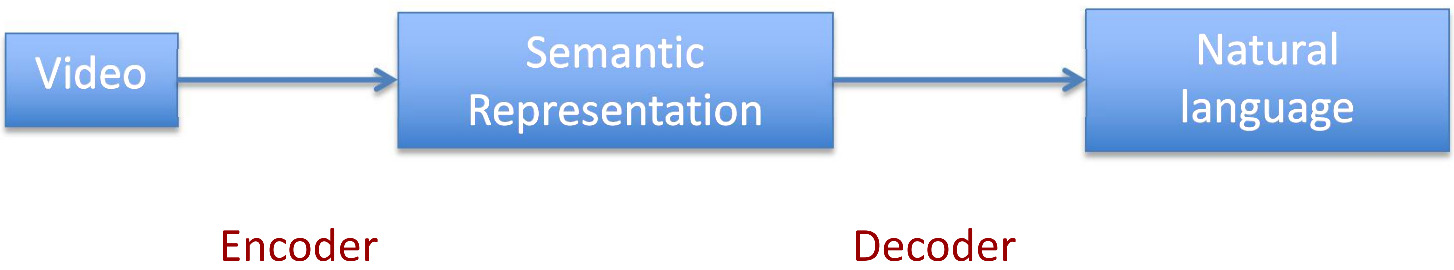
- For each test video, retrieve the 3 most similar videos according to the SVO triplet
- Ask workers to rate, on a scale of 1 to 5, how relevant the retrieved videos are with respect to the given video.
- Average rating obtained is **1.99**

Related works

Early works

- **Translating video content to natural language descriptions [10]**

Encoder-decoder framework:
Video description is phrased as a translation problem from video content to natural language and used a semantic representation of the video content as intermediate step.



Encoder : Conditional Random Field
Decoder : Statistical Machine Translation

[10] Rohrbach, Marcus, et al. "Translating video content to natural language descriptions." Proceedings of the IEEE International Conference on Computer Vision. 2013.

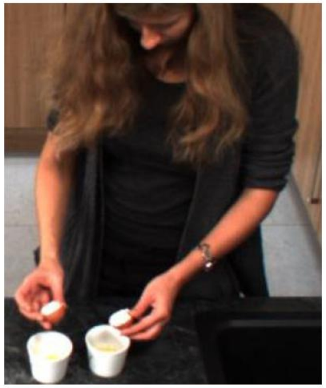
Related works

Early works

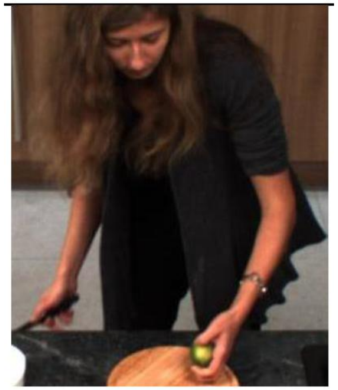
- **Translating video content to natural language descriptions**

Experimental results on TACoS

CRF+SMT: the person cracks the eggs
Human: the person dumps any remaining whites of the eggs from the shells into the cup with the egg whites



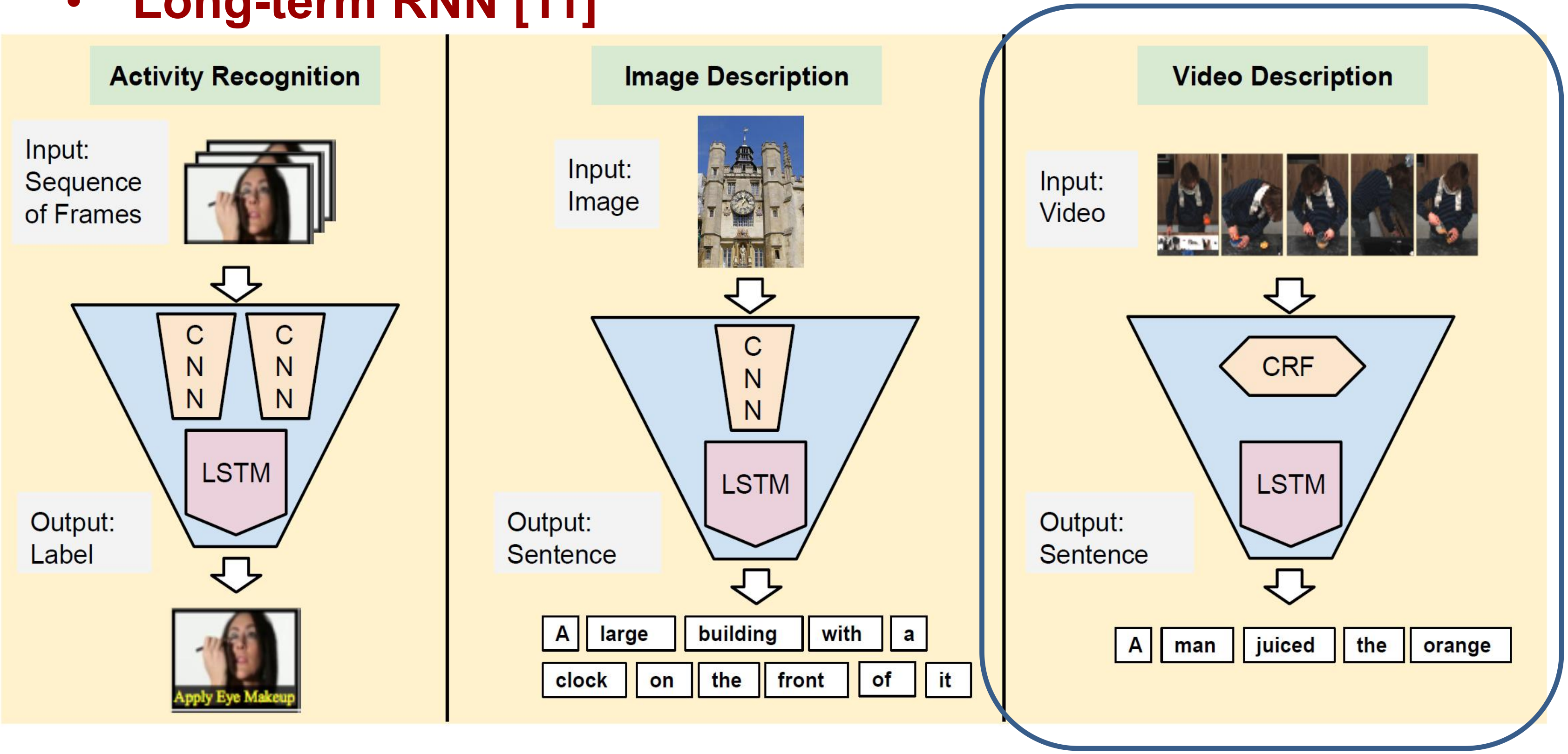
CRF+SMT: the person gets out a cutting board from the loaf of bread from the fridge
Human: the person gets the lime, a knife and a cutting board



Related works

Recent works

- **Long-term RNN [11]**



[11] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

Related works

Recent works

- **Long-term RNN**

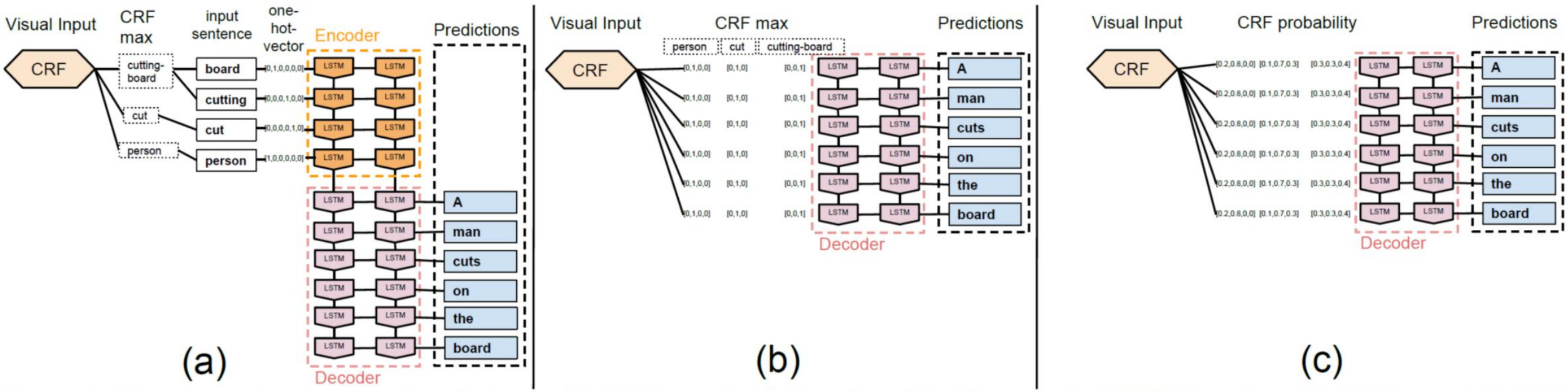


Figure 4: Our approaches to video description. (a) LSTM encoder & decoder with CRF max (b) LSTM decoder with CRF max (c) LSTM decoder with CRF probabilities. (For larger figure zoom or see supplemental).

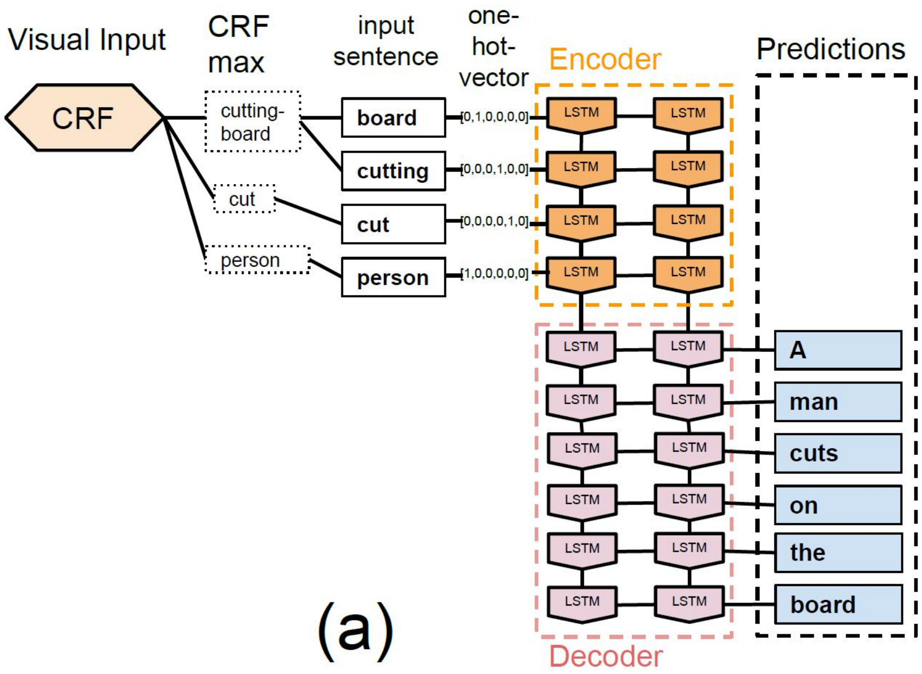
Related works

Recent works

- **Long-term RNN**

LSTM as both encoder and decoder

Use CRF max



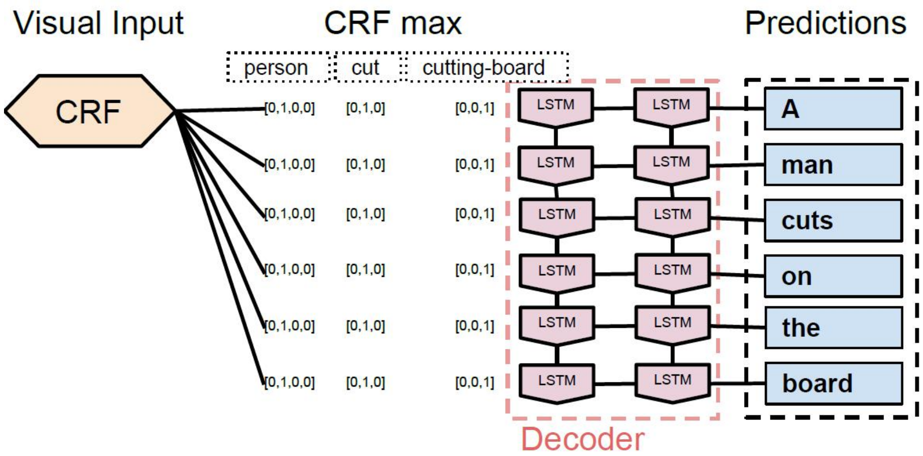
Related works

Recent works

- **Long-term RNN**

LSTM as decoder

Use CRF max



(b)

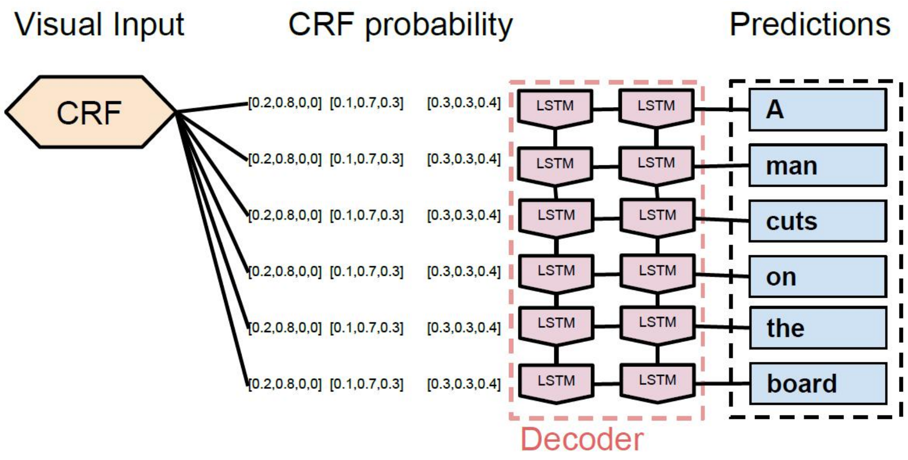
Related works

Recent works

- **Long-term RNN**

LSTM as decoder

Use CRF probabilities



(c)

Related works

Recent works

- **Long-term RNN**

Experimental results on TACoS

Architecture	Input	BLEU (%)
SMT[9]	CRF max	24.9
LSTM (a)	CRF max	25.3
LSTM (b)	CRF max	27.4
LSTM (c)	CRF probilities	**28.8**

Related works

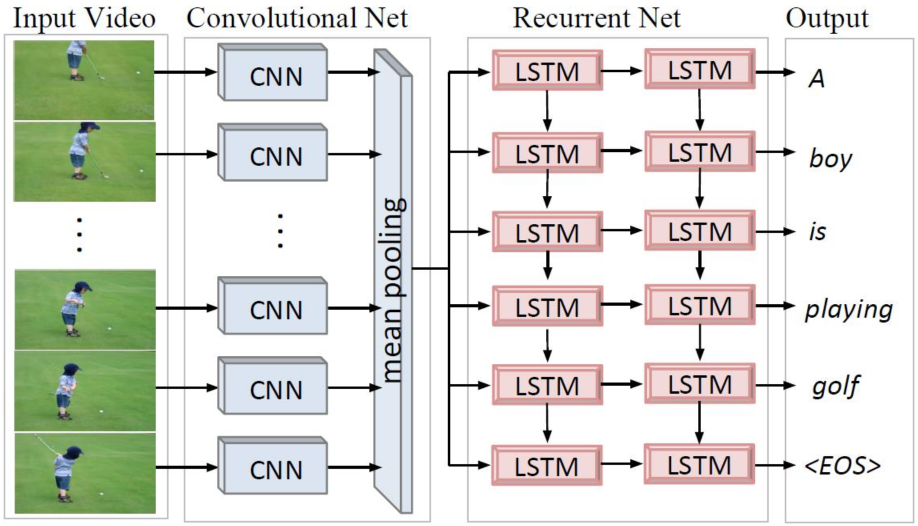
Recent works

- **Mean pooling [12]**

Basic encoder-decoder framework

Encoder: pre-trained CNN
for each frame separately
mean-pooling on all frames

Decoder: LSTM



[15] Venugopalan, Subhashini, et al. "Translating videos to natural language using deep recurrent neural networks." arXiv preprint arXiv:1412.4729 (2014).

Related works

Recent works

- **Mean pooling [12]**

Experimental results using METEOR (%)

Methods	MSVD	MVAD	MPII-MD
Mean pool - AlexNet	26.9		
Mean pool - VGG	27.7	6.1	6.7
Mean pool - AlexNet COCO pre-trained	29.1		
Mean pool - GoogleNet	28.7		

Related works

Recent works

- **Temporal attention [13]**

Basic encoder-decoder framework

Encoder: pre-trained CNN on ImageNet
used for each frame separately
+ temporal information

Decoder: LSTM

[13] Yao, Li, et al. "Describing videos by exploiting temporal structure." Proceedings of the IEEE International Conference on Computer Vision. 2015.

Related works

Recent works

- **Temporal attention**

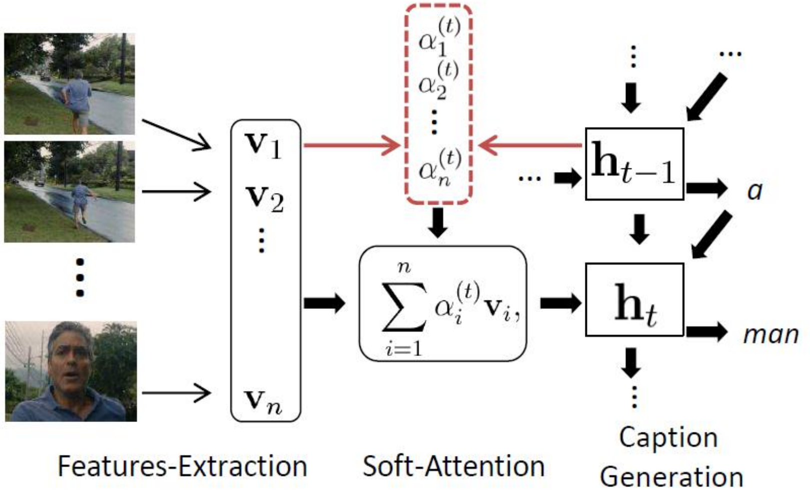
Exploiting temporal structure

Local: 3D-CNN

three 3D convolutional layer

temporal features obtained by max-pooling

Global: temporal attention mechanism



Related works

Recent works

- **Temporal attention**

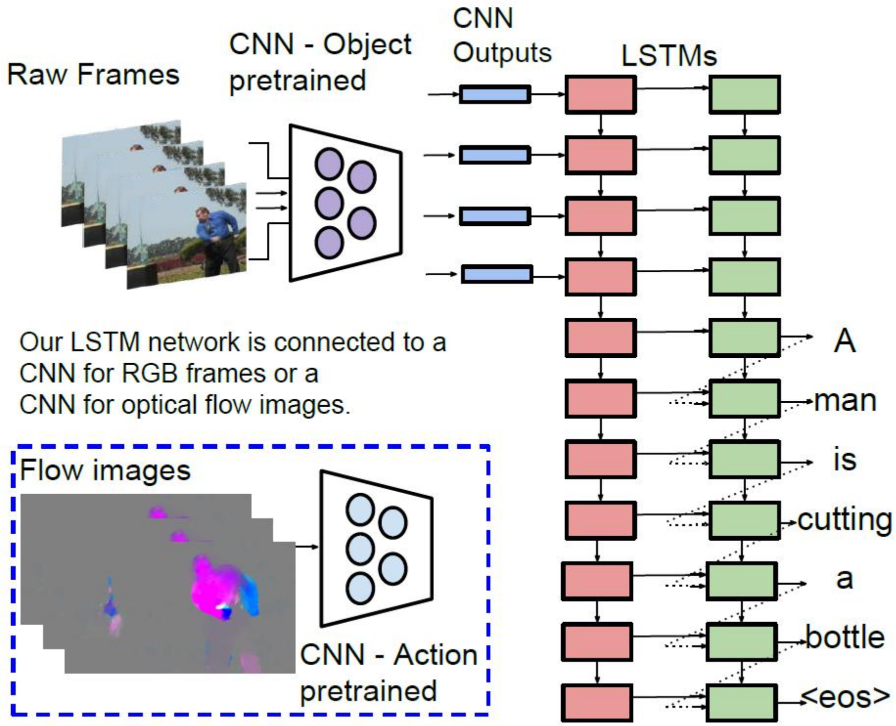
  Experimental results using METEOR (%)

Methods	MSVD	MVAD	MPII-MD
Mean pool - GoogleNet	28.7		
Temporal attention - GoogleNet	29.0		
Temporal attention - GoogleNet + 3D-CNN	29.6	4.3	

Related works

Recent works

- **S2VT [14]**



[14] Venugopalan, Subhashini, et al. "Sequence to sequence-video to text." Proceedings of the IEEE International Conference on Computer Vision. 2015.

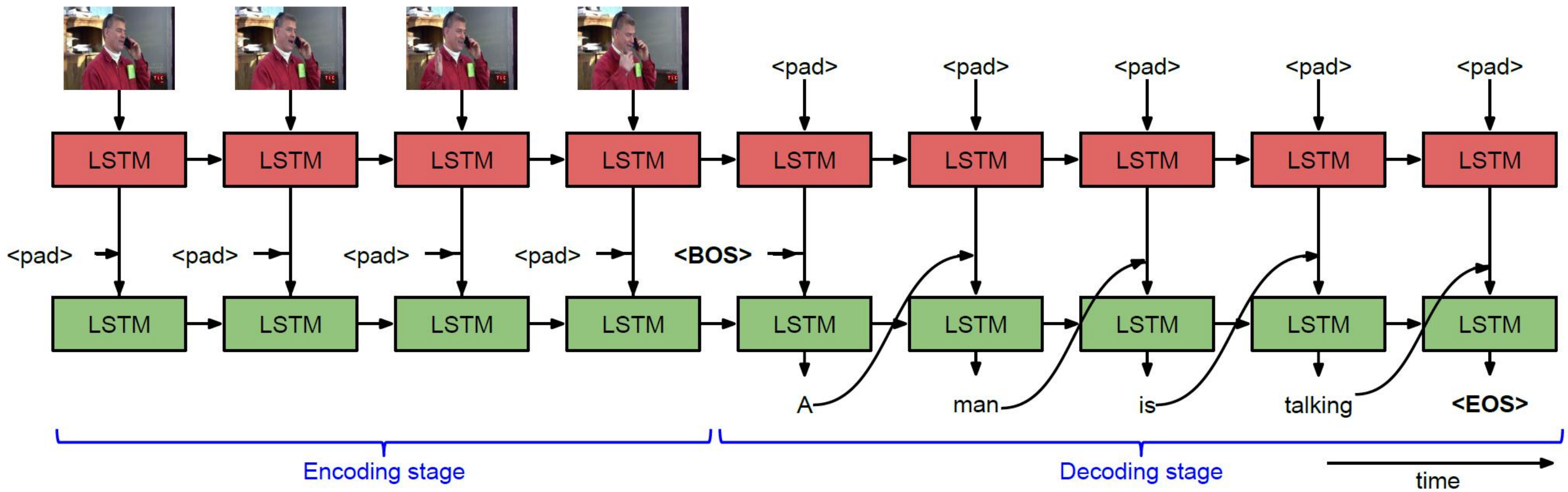
Related works

Recent works

- **S2VT [17]**

No separate encoder-decoder

Use the same LSTM for both encoder and decoder



Related works

Recent works

- **S2VT [17]**

Experimental results using METEOR (%)

Methods	MSVD	MVAD	MPII-MD
Mean pool - AlexNet	26.9		
Mean pool - VGG	27.7	6.1	6.7
Mean pool - GoogleNet	28.7		
Temporal attention - GoogleNet	29.0		
Temporal attention - GoogleNet + 3D-CNN	29.6	4.3	
S2VT (Flow) - AlexNet	24.3		
S2VT (RGB) - AlexNet	27.9		
S2VT (RGB) - VGG	29.2	6.7	7.1
S2VT (RGB + Flow) - VGG for RGB, AlexNet for Flow	29.8		

Related works

Recent works

- **hRNN [7]**

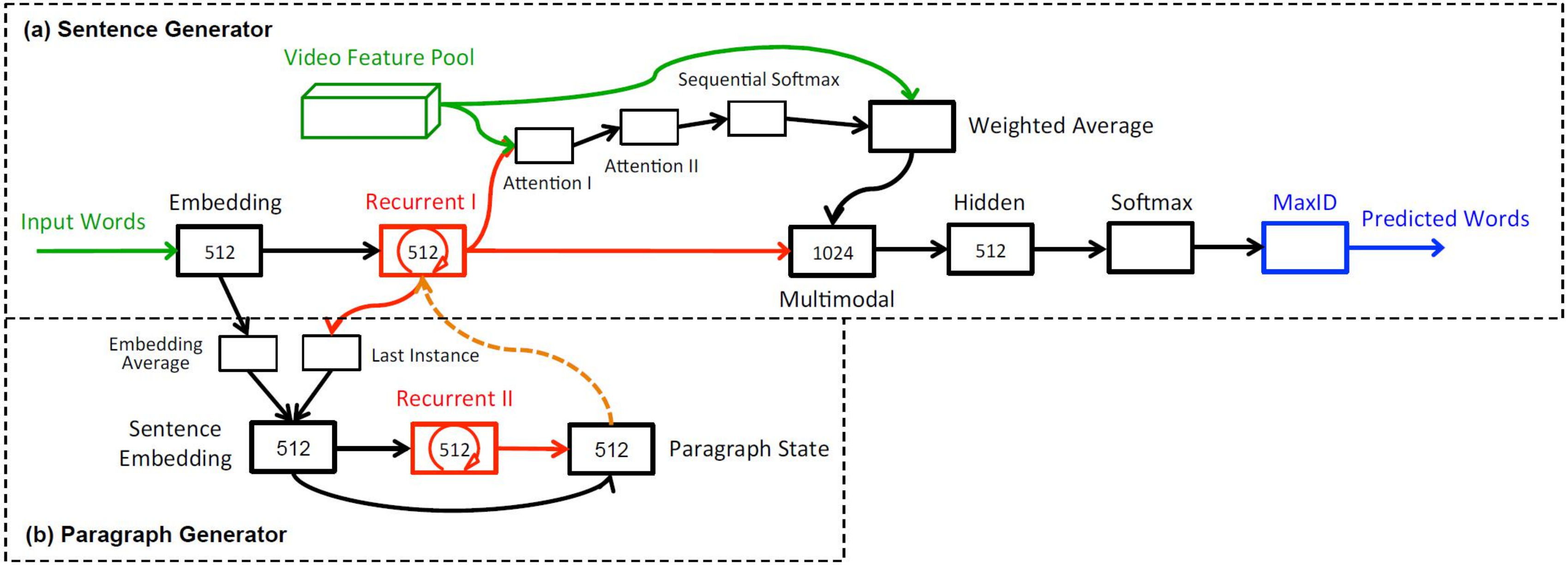


Figure 2. Our hierarchical RNN for video captioning. Green denotes the input to the framework, blue denotes the output, and red denotes the recurrent components. The orange arrow represents the reinitialization of the sentence generator with the current paragraph state. For simplicity, we only draw a single video feature pool in the figure. In fact, both appearance and action features go through a similar attention process before they are fed into the multimodal layer.

Related works

Recent works

- **hRNN**

Two language generators: sentence generator and paragraph generator

Multimodal layer after the recurrent layer to combine video content features

2D CNN for frame feature extraction, 3D CNN for video feature extraction

Related works

Recent works

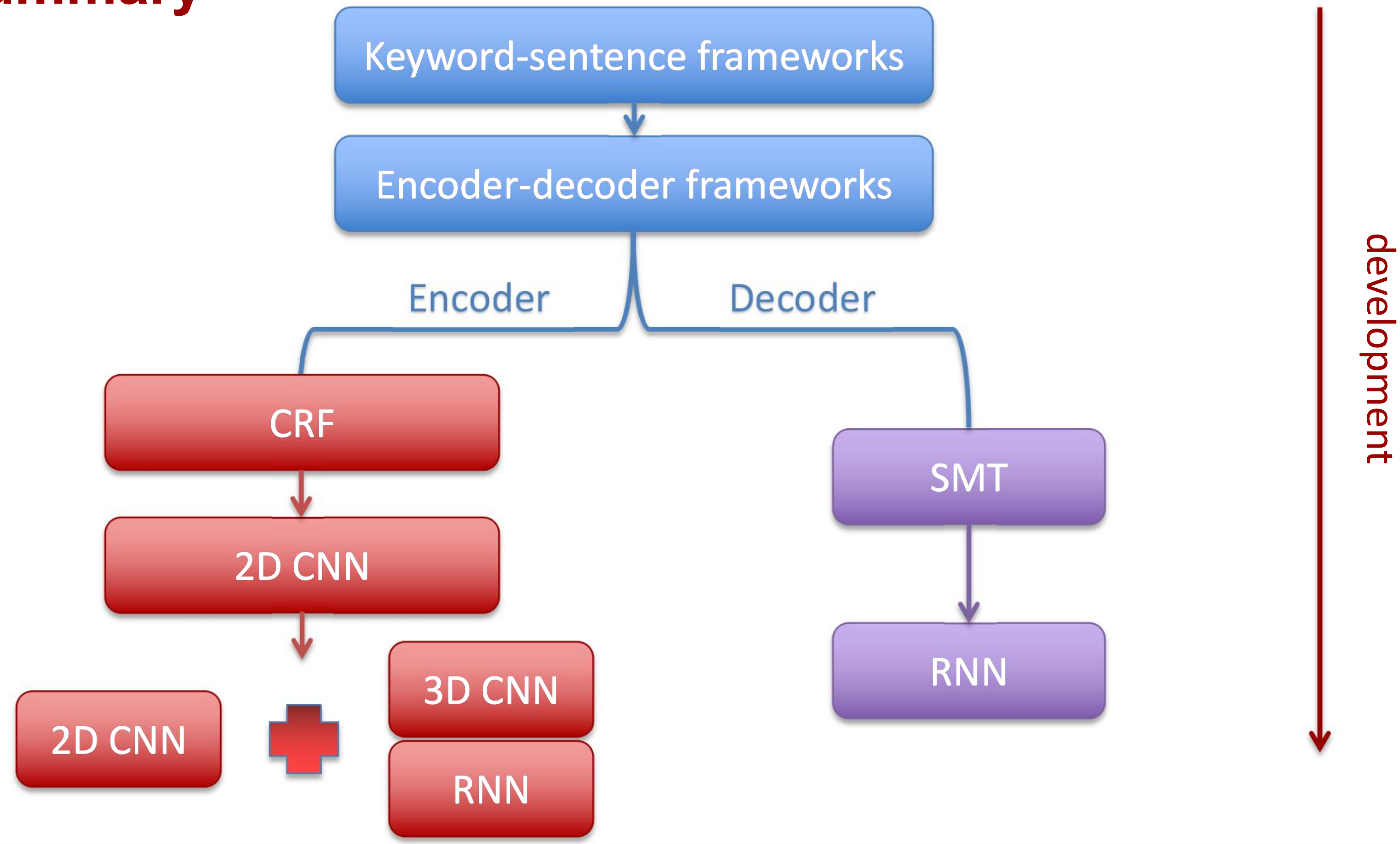
- **hRNN**

  Experimental results using METEOR (%)

Methods	MSVD
Mean pool - VGG	27.7
Temporal attention - GoogleNet + 3D-CNN	29.6
S2VT (RGB) - VGG	29.2
S2VT (RGB + Flow) - VGG for RGB, AlexNet for Flow	29.8
hRNN - VGG	31.1
hRNN- C3D	30.3
hRNN - VGG + C3D	32.6

Related works

Summary



Related works

Future

- **Encoder-decoder framework**
  - encoder: +scene classification

  - encoder to decoder: better structure

  - decoder

- **Other framework**

Conclusion

Video description is ...

- **important**
  tagging
  indexing
  human-robot interaction

- **difficult**
  implementation
  evaluation

- **under development**
  datasets
  evaluation methods
  algorithms