Video Paragraph Captioning using Hierarchical Recurrent Neural Networks

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, Wei Xu

Problem

Given a video, generate a paragraph (multiple sentences).

Problem

Given a video, generate a paragraph (multiple sentences).

The person entered the kitchen.
The person opened the drawer.
The person took out a knife and a sharpener.
The person sharpened the knife.
The person cleaned the knife.

Problem

Given a video, generate a paragraph (multiple sentences).

The person entered the kitchen.
The person opened the drawer.
The person took out a knife and a sharpener.
The person sharpened the knife.
The person cleaned the knife.

VS.

The person sharpened the knife in the kitchen.

Motivation

Inter-sentence dependency (semantics context)

Motivation

Inter-sentence dependency (semantics context)

The person took out some potatoes.

Motivation

Inter-sentence dependency (semantics context)

The person took out some potatoes.

The person peeled the potatoes.

The person turned on the stove.

Motivation

Inter-sentence dependency (semantics context)

The person took out some potatoes.

The person peeled the potatoes.

The person turned on the stove.

We want to model this dependency.

Hierarchy

A paragraph is inherently hierarchical.

Hierarchy

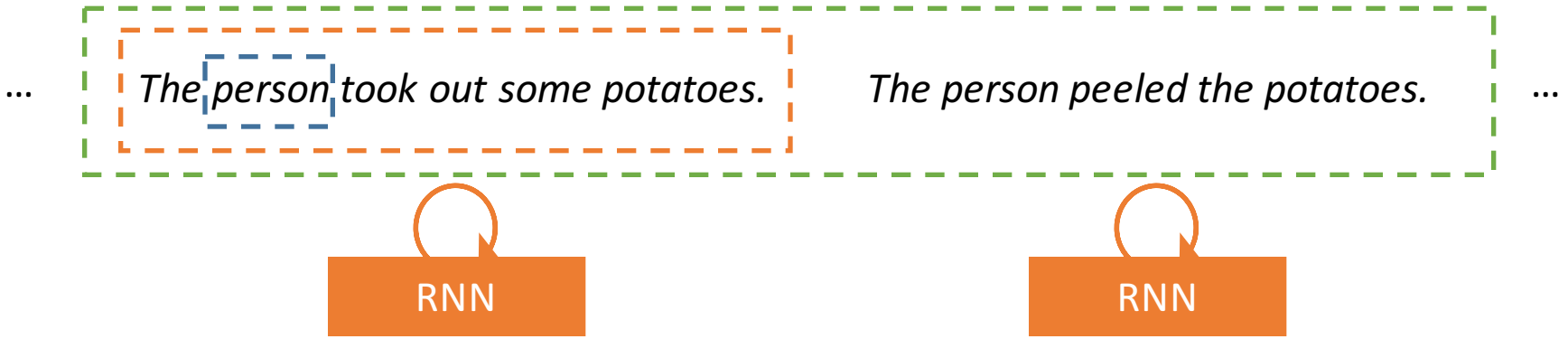
A paragraph is inherently hierarchical.

Hierarchy

A paragraph is inherently hierarchical.

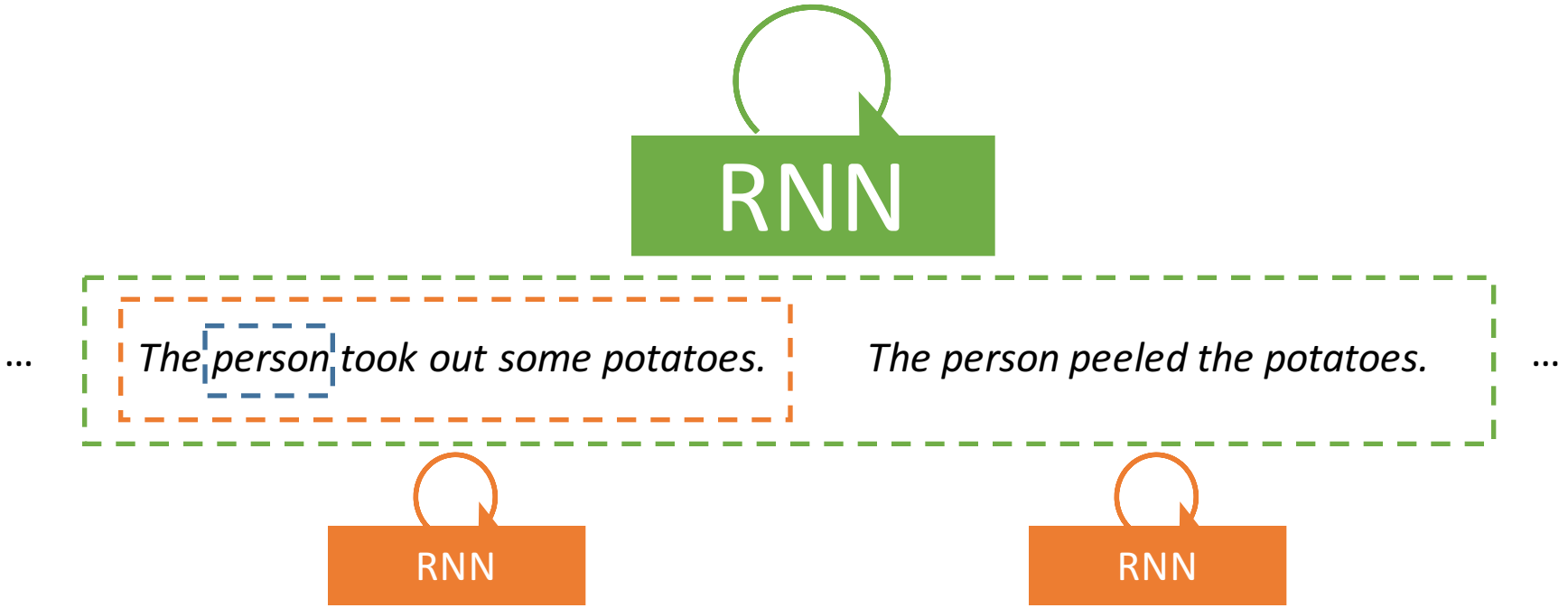
Hierarchy

A paragraph is inherently hierarchical.

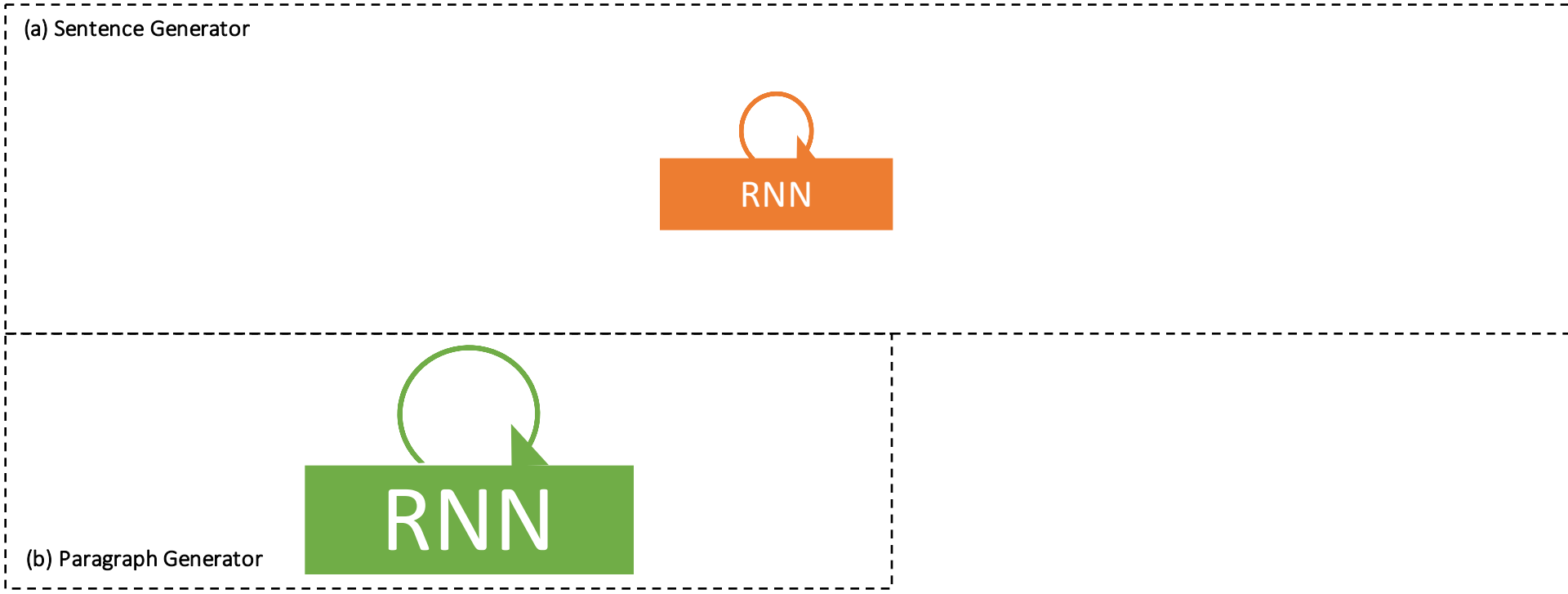


Hierarchy

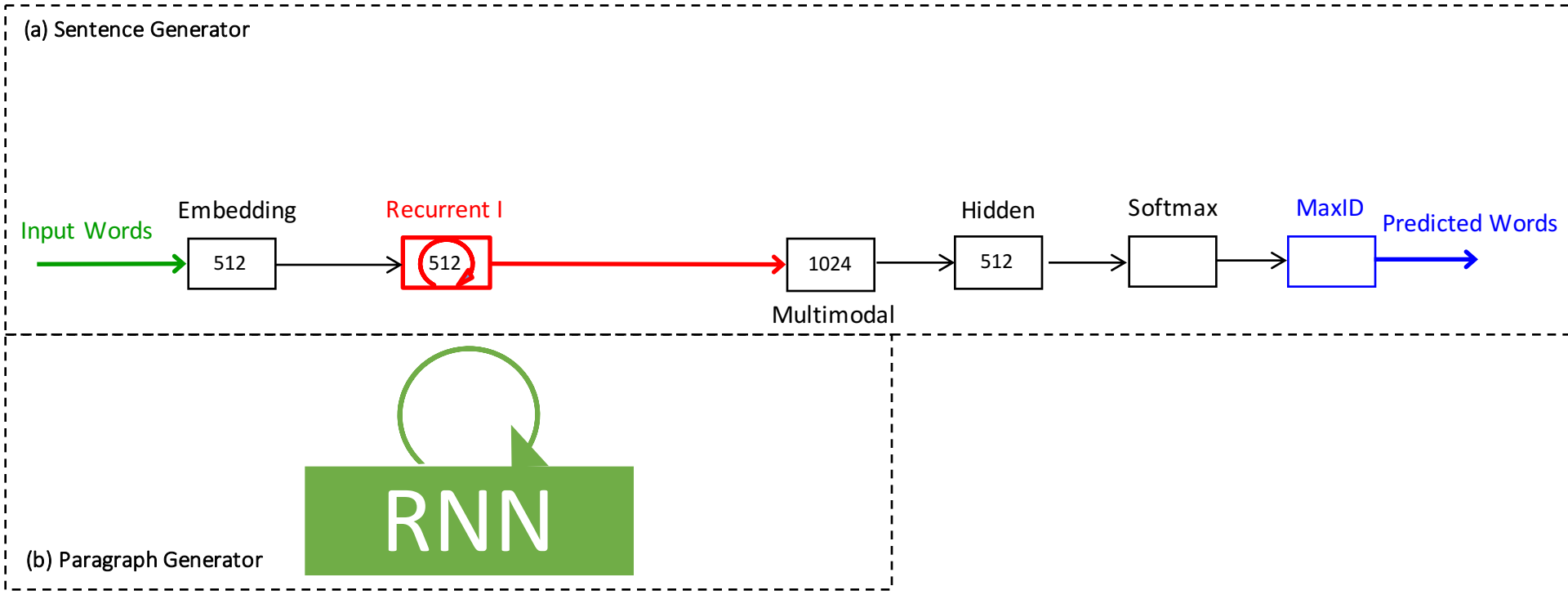
A paragraph is inherently hierarchical.



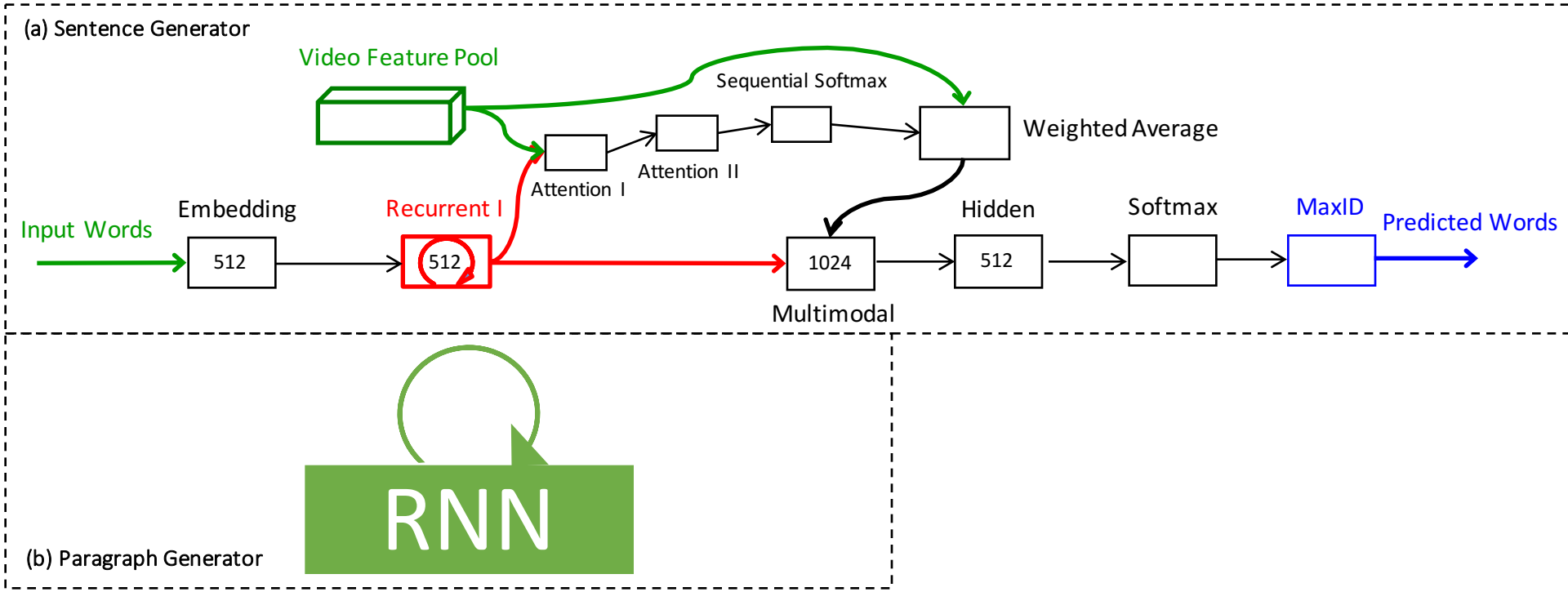
Framework



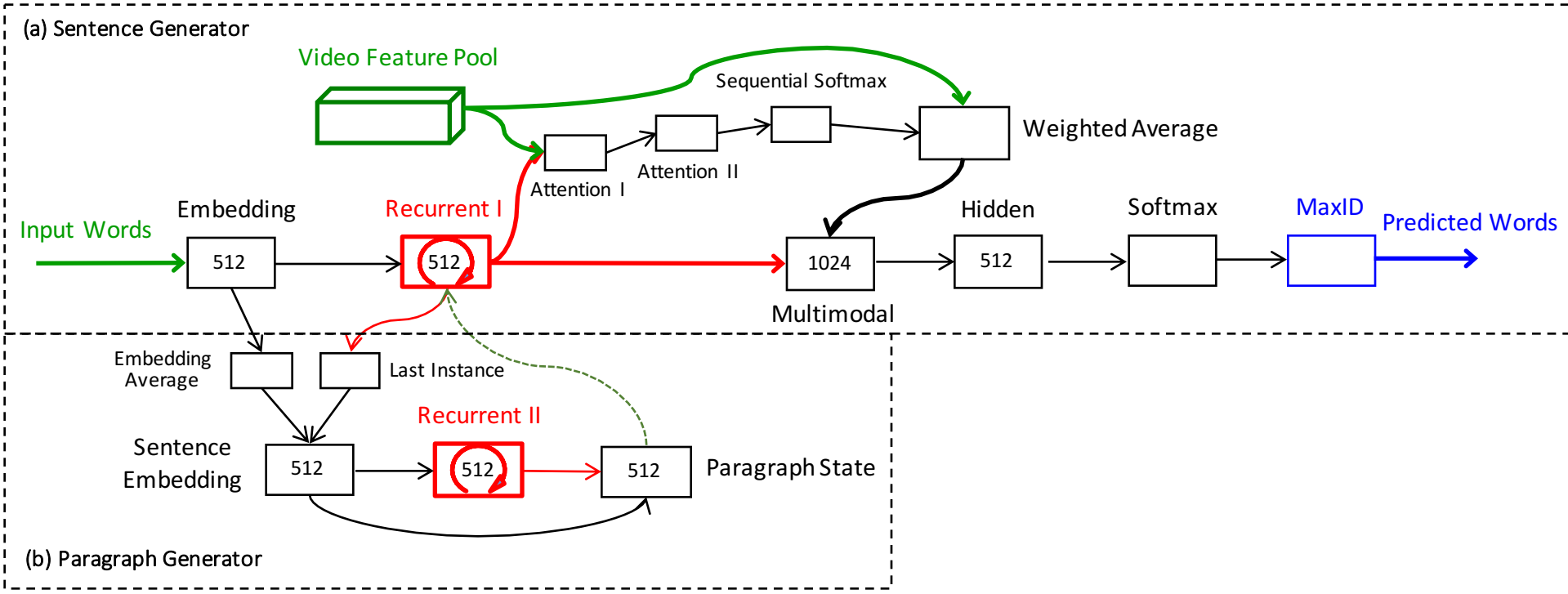
Framework – language model



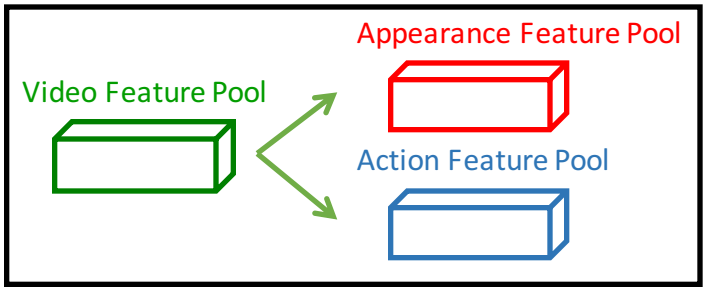
Framework – attention model for video feature



Framework – paragraph model



Visual Features



Object appearance:

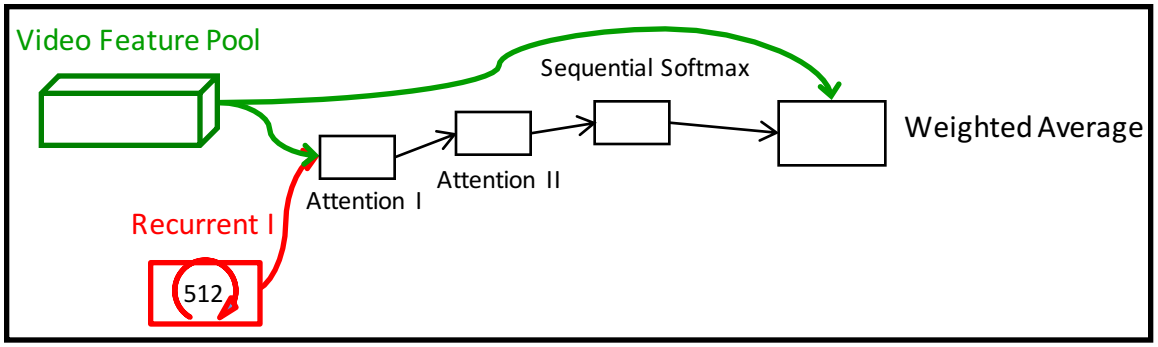
VGG-16 (fc7) [Simonyan *et al.*, 2015], pre-trained on ImageNet dataset

Action:

C3D (fc6) [Tran *et al.*, 2015], pre-trained on Sports-1M dataset

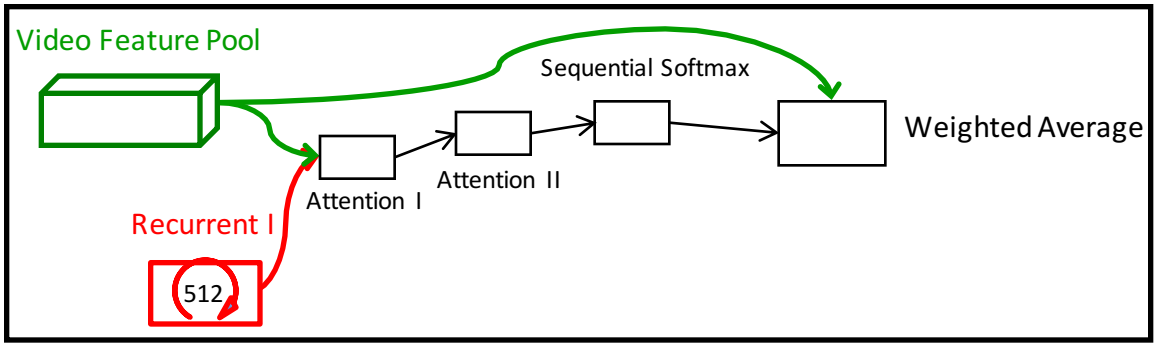
Dense Trajectories+Fisher Vector [Wang *et al.*, 2011]

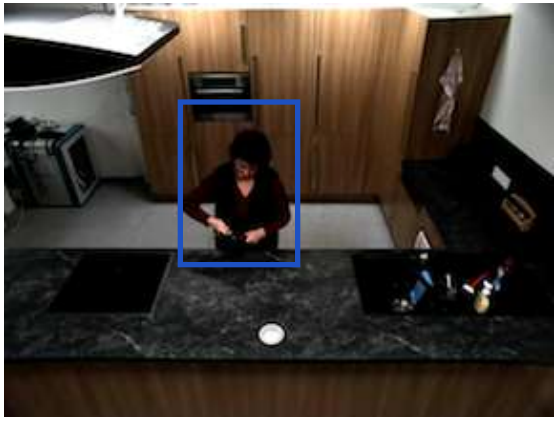
Attention Model

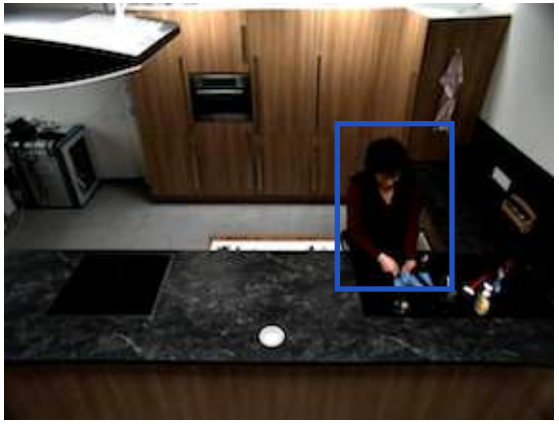


Learning spatial & temporal attention simultaneously

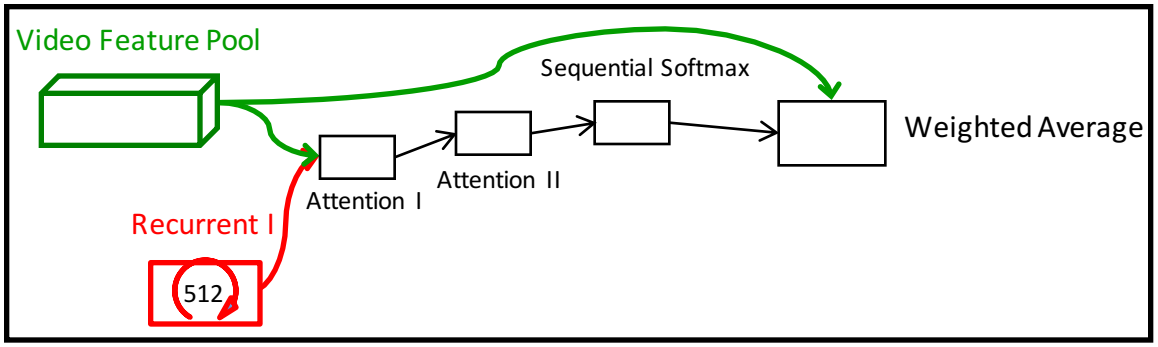
Attention Model

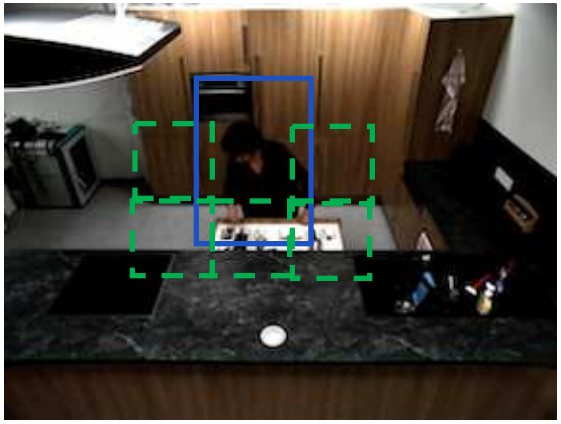


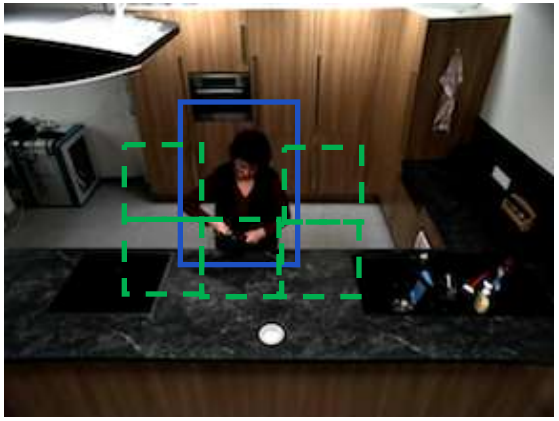


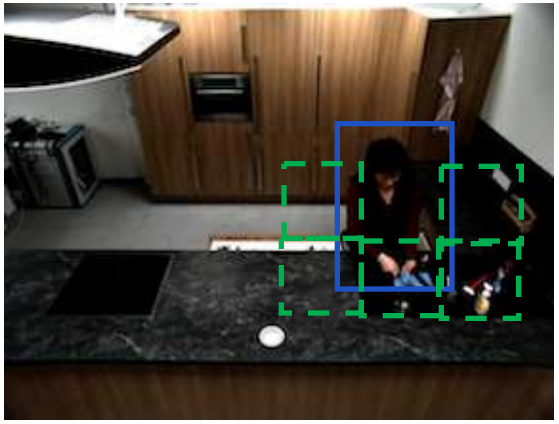


Attention Model

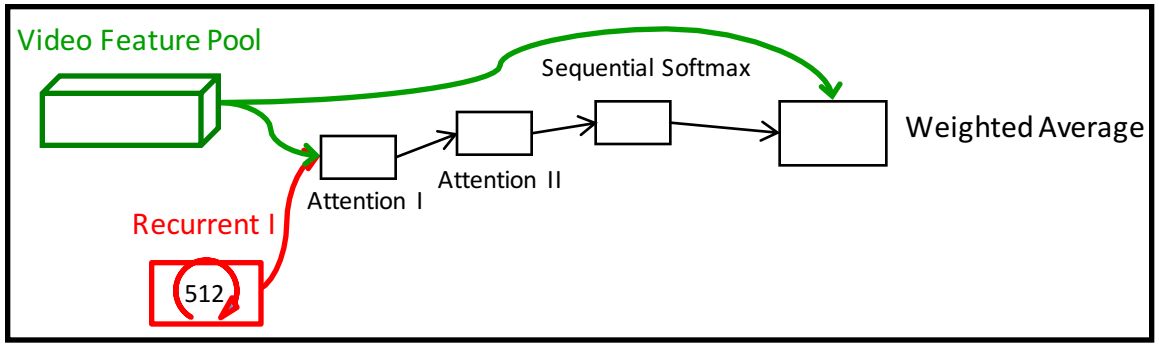


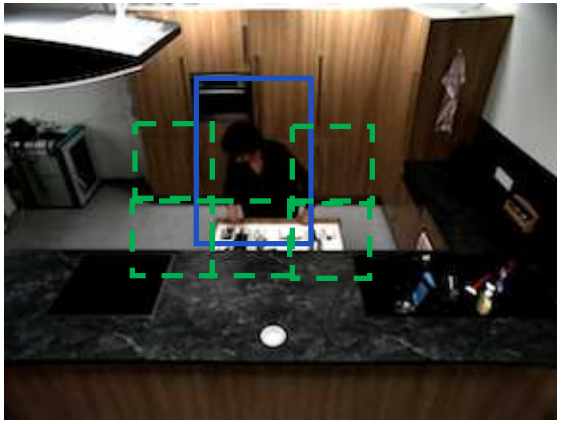


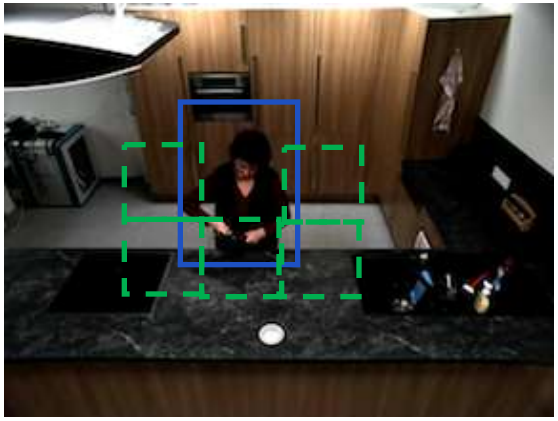


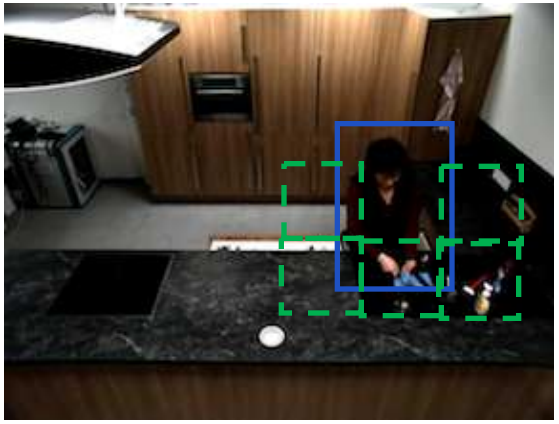


Attention Model



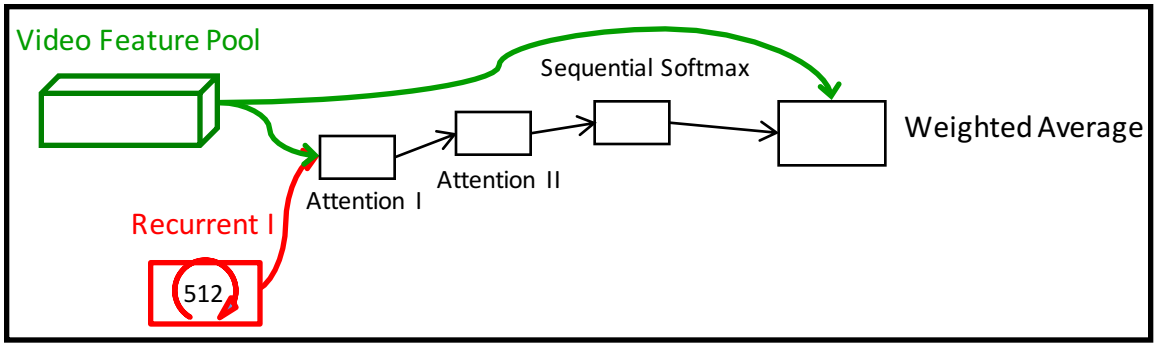


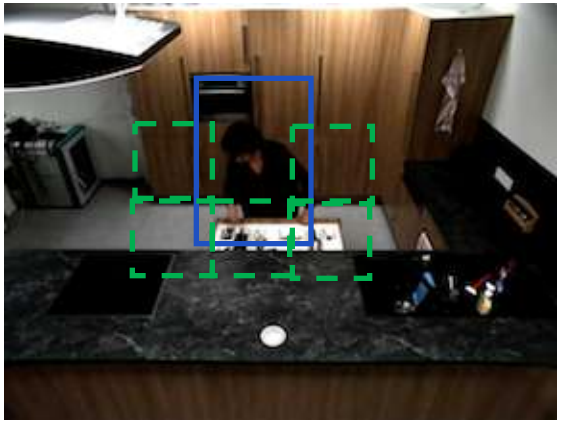


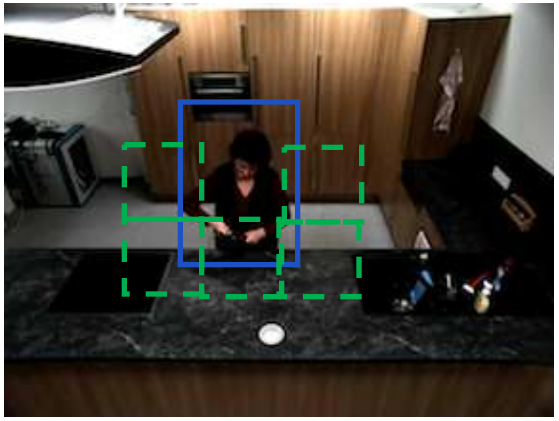


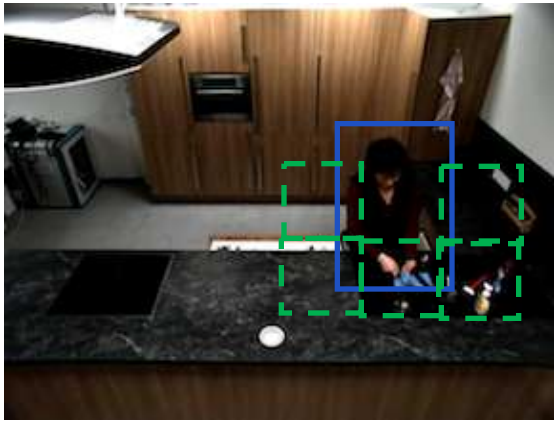
feature pool ... i-1 i i+1 ...

Attention Model





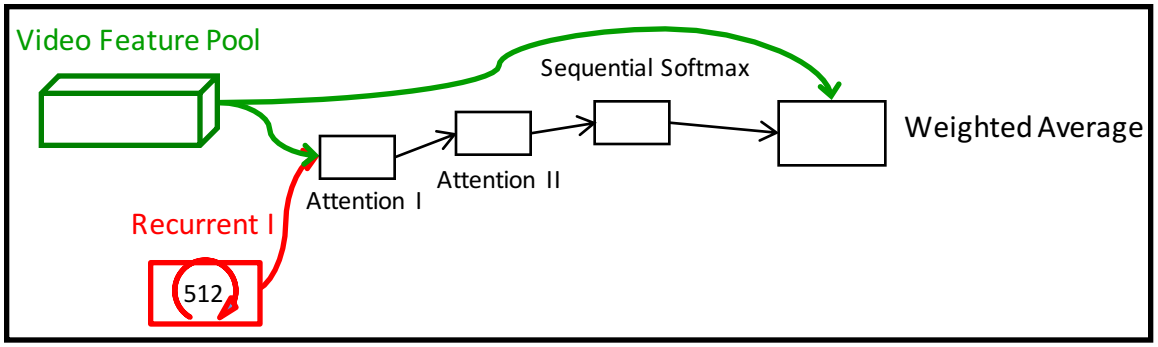


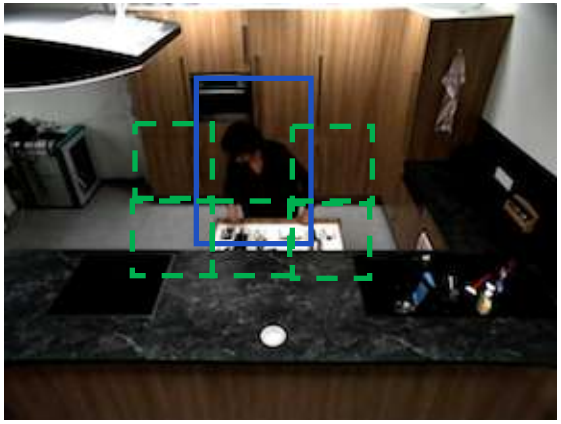


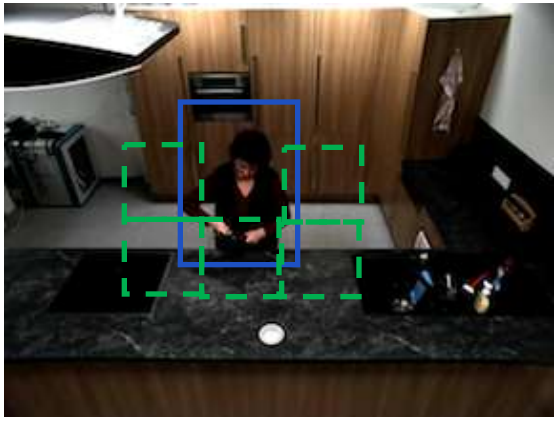
feature pool ... i-1 i i+1 ...

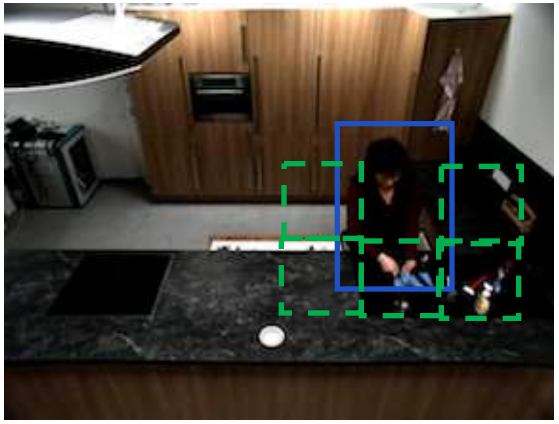
previous recurrent state t-1

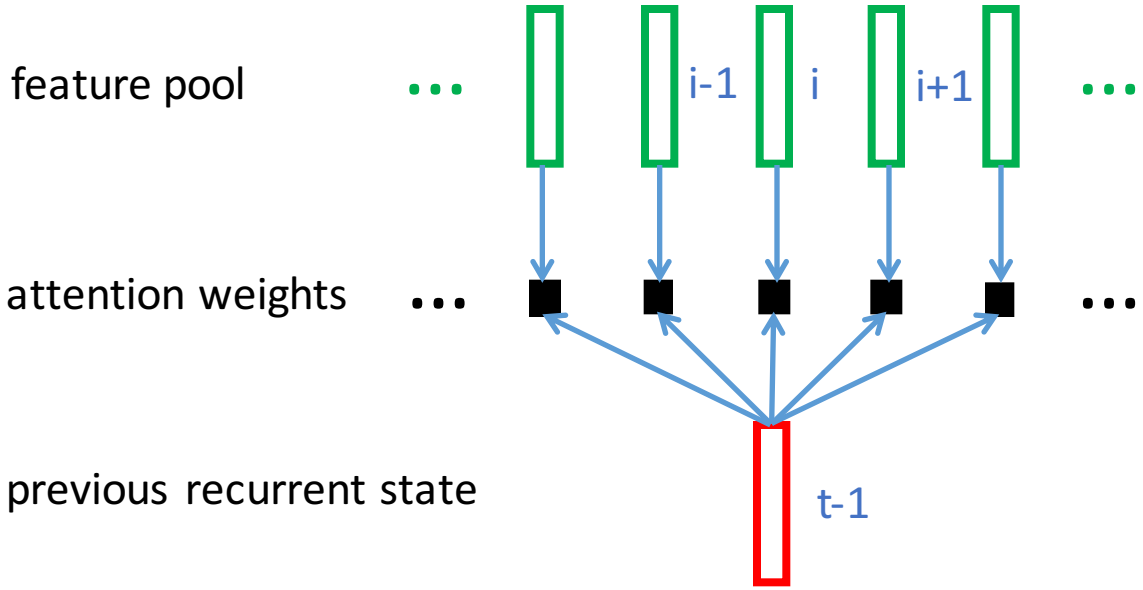
Attention Model



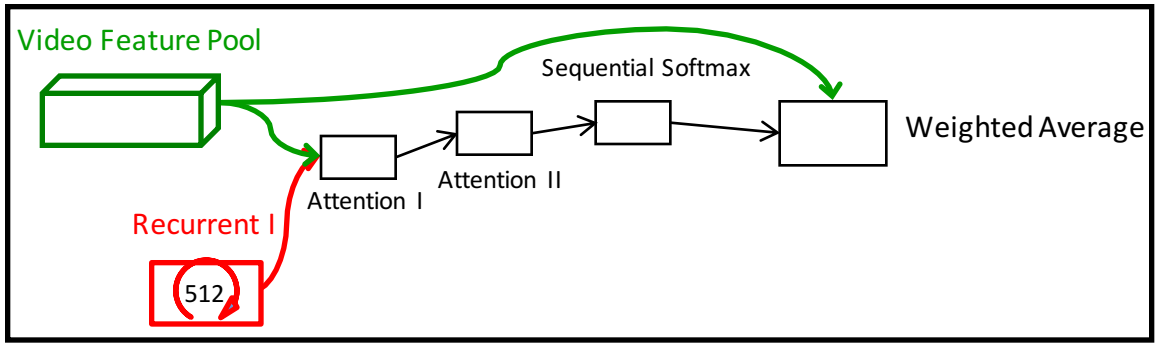


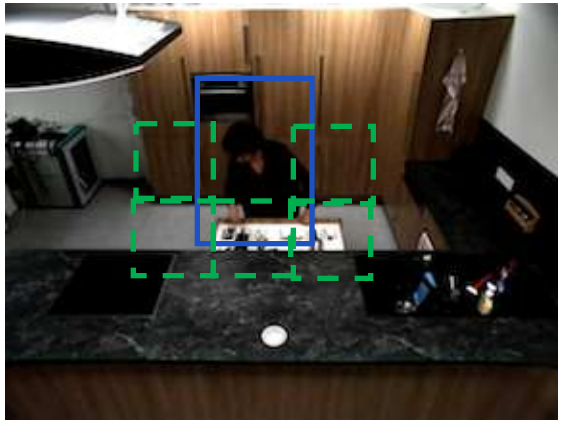


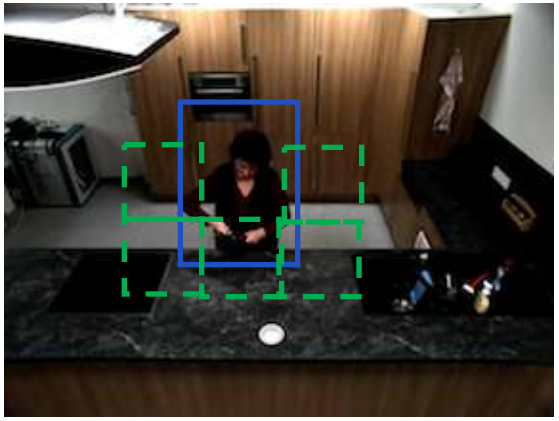


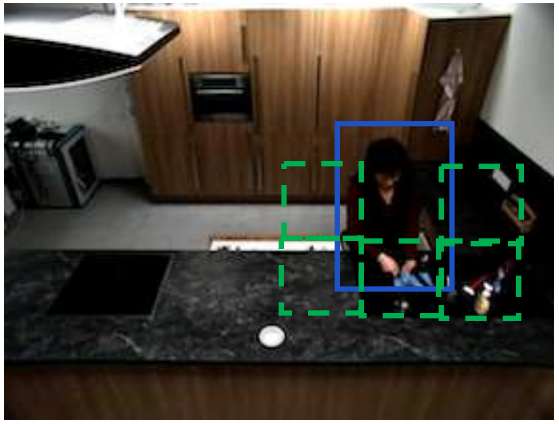


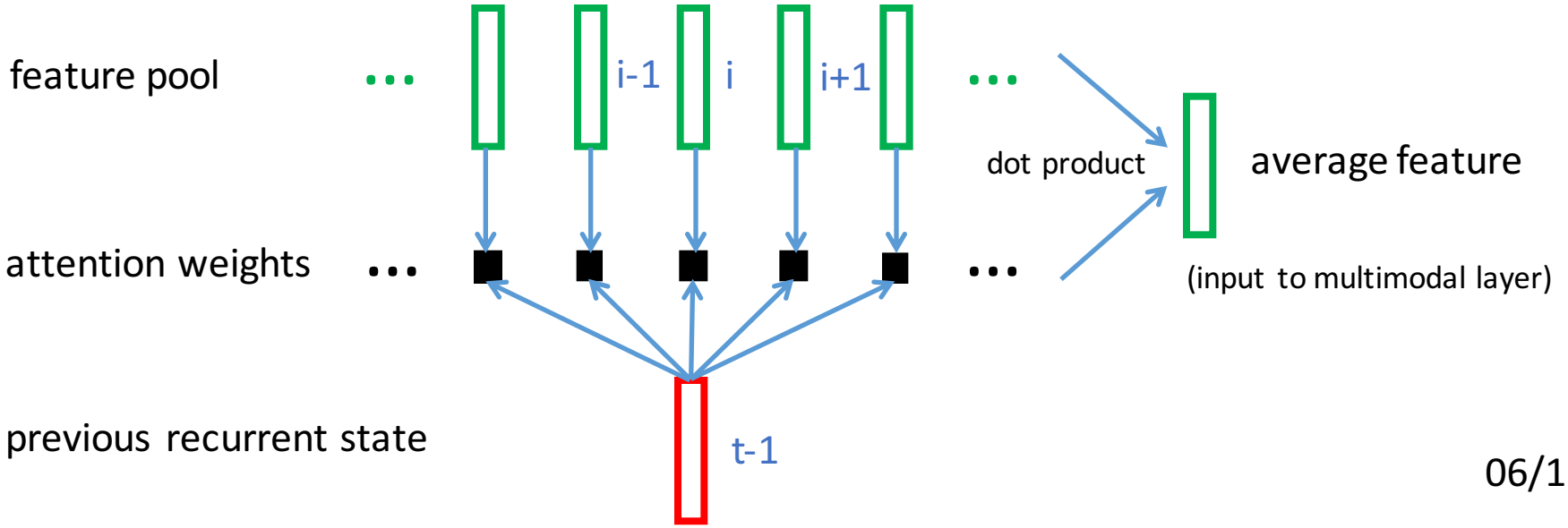
Attention Model



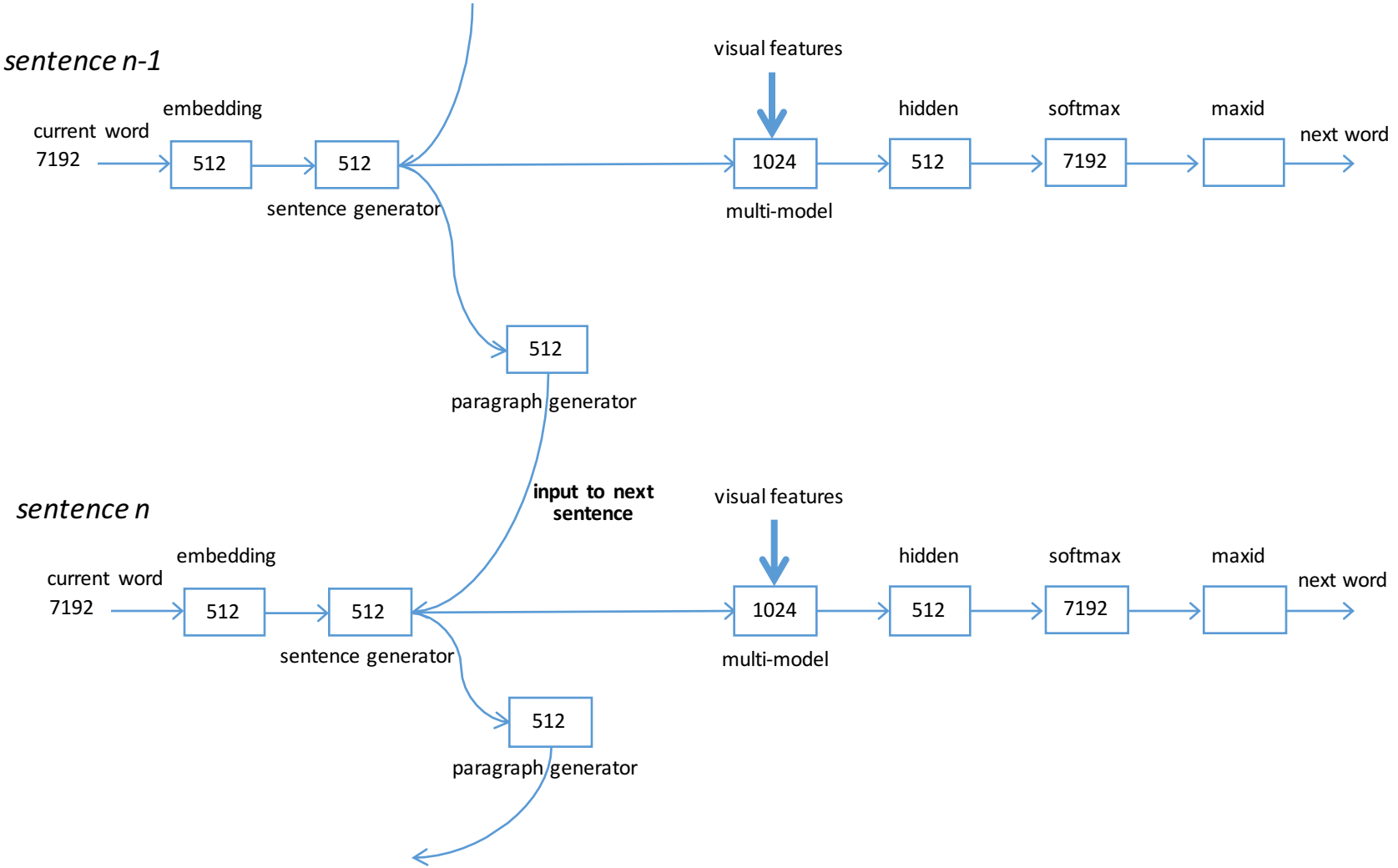




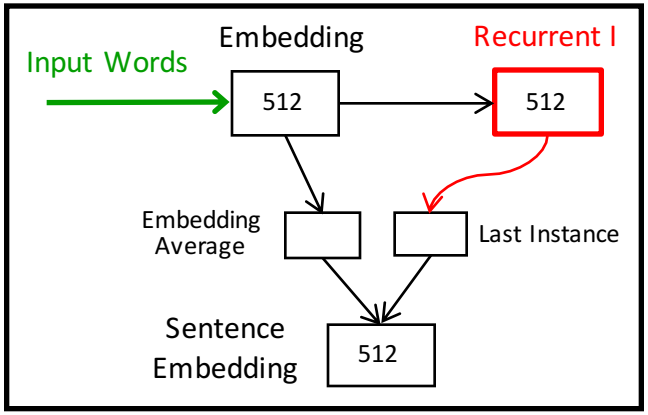


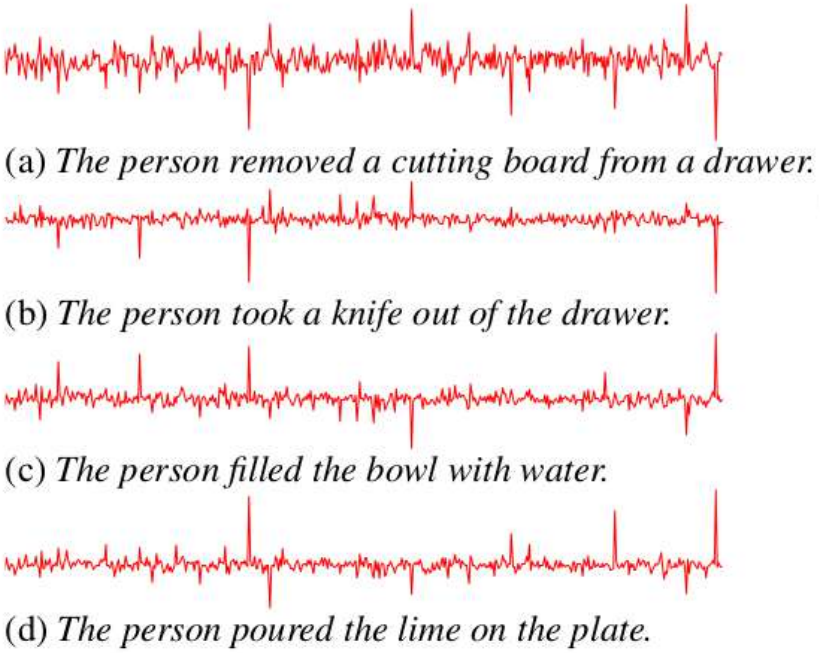


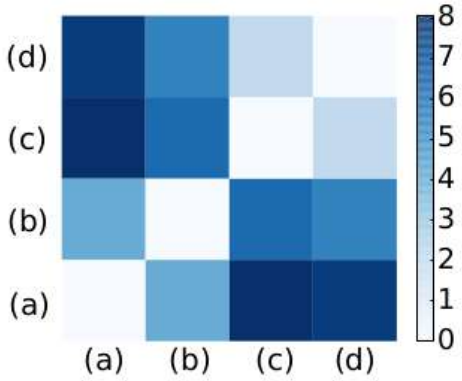
Paragraph Generator Unrolled



Sentence Embedding







Experiments - Setup

Two datasets:

YouTube2Text

> open-domain

> 1,970 videos, ~80k video-sentence pairs, 12k unique words

> only one sentence for a video (*special case*)

TACoS-MultiLevel

> closed-domain: cooking

> 173 videos, 16,145 intervals, ~40k interval-sentence pairs, 2k unique words

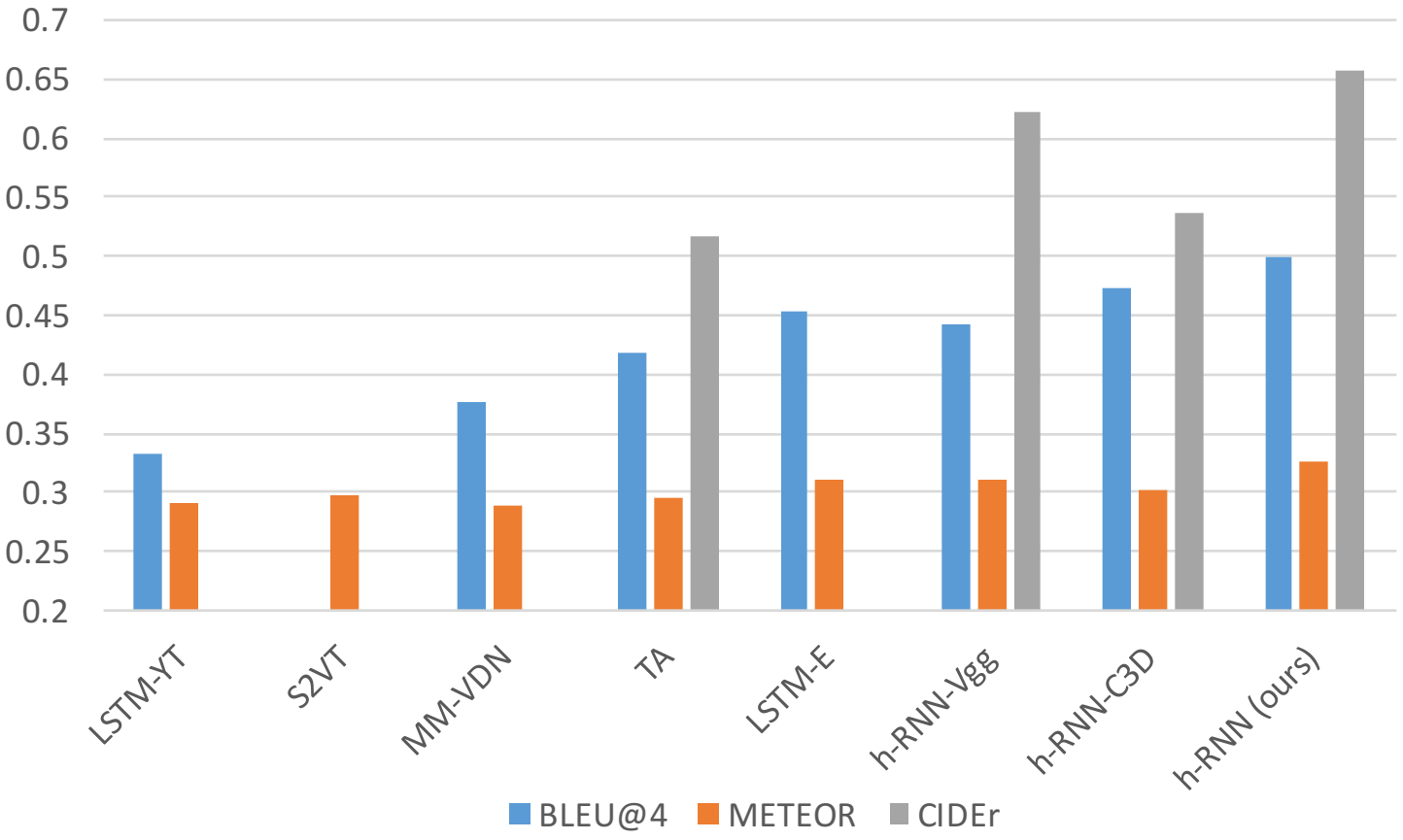
> several dependent sentences for a video

Three evaluation metrics:

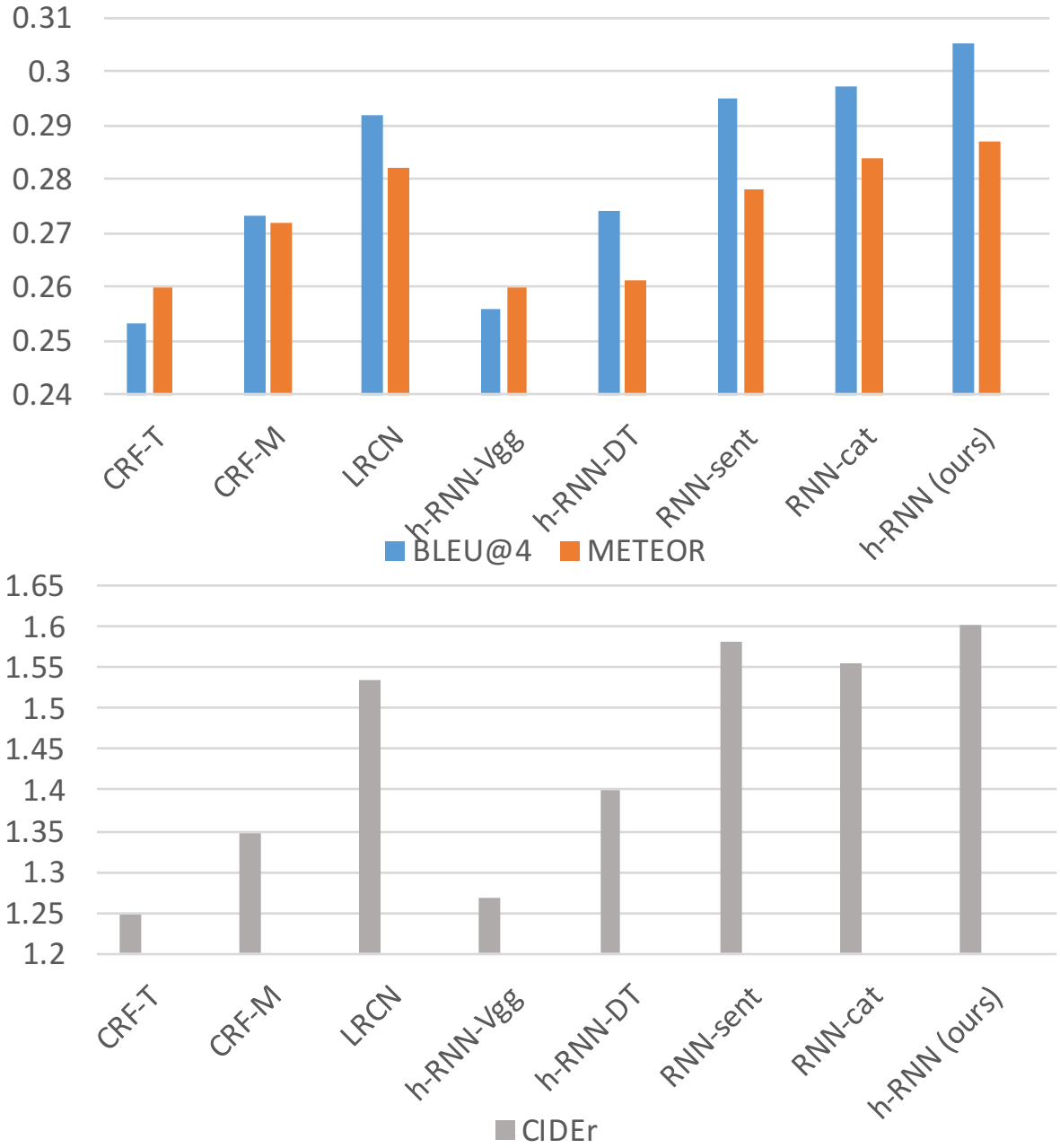
BLEU	[Papineni *et al.,* 2002]
METEOR	[Banerjee and Lavie, 2005]
CIDEr	[Vedantam *et al.,* 2015]

The higher, the better.

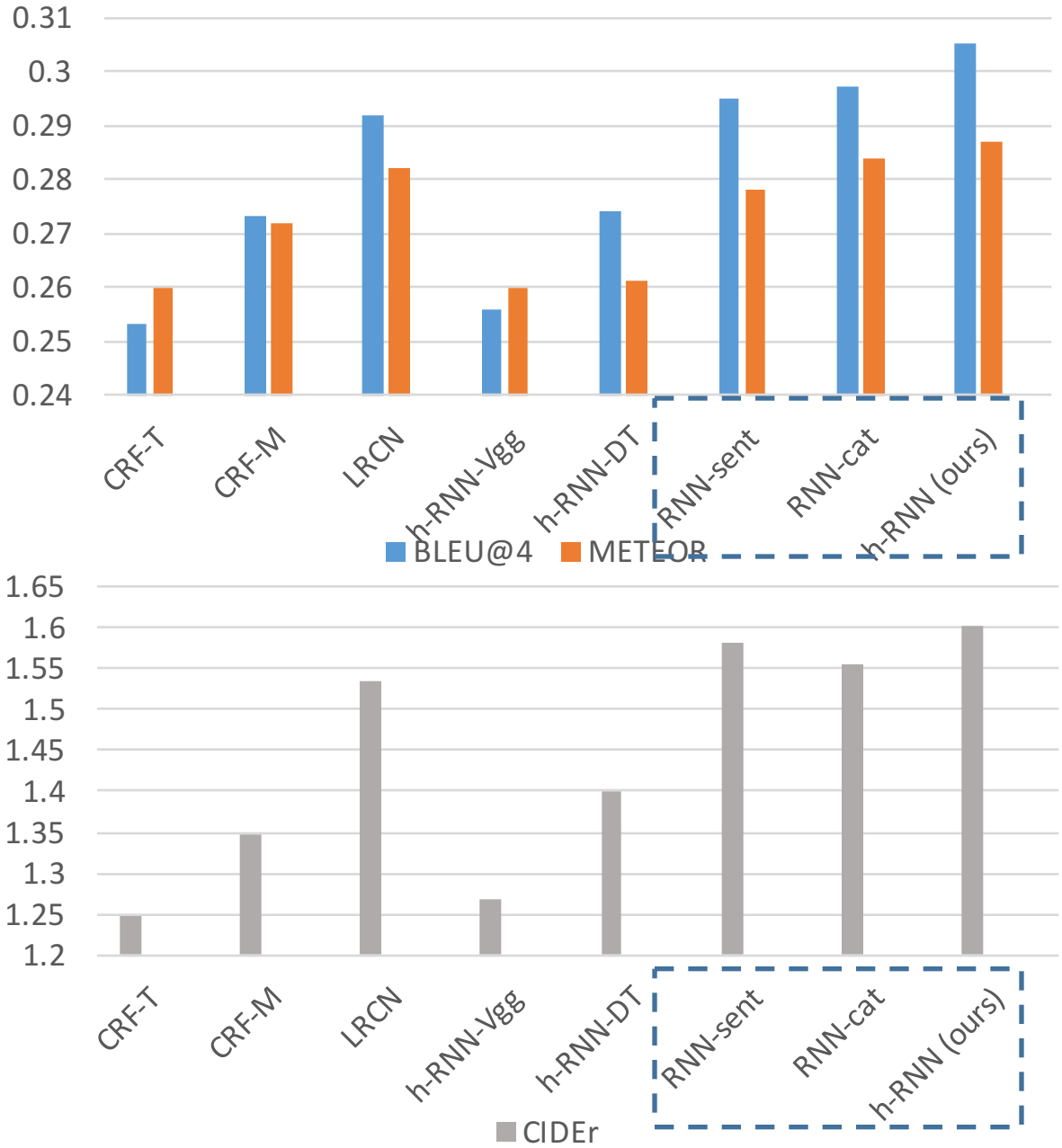
Experiments - YouTube2Text



Experiments - TACoS-MultiLevel

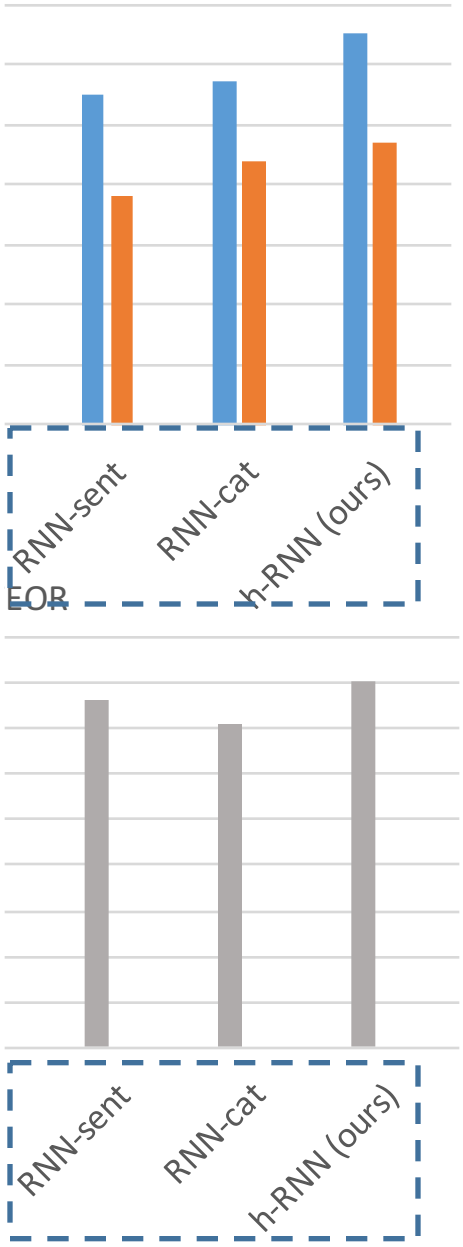


Experiments - TACoS-MultiLevel



Experiments - TACoS-MultiLevel

Evaluation metric scores are not always reliable, we need further comparison.

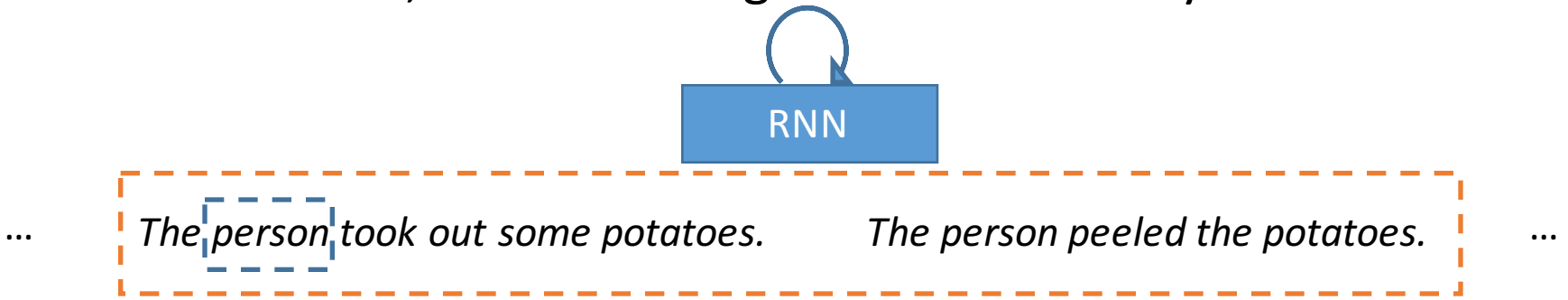


RNN-cat vs. h-RNN

RNN-cat vs. h-RNN

RNN-cat

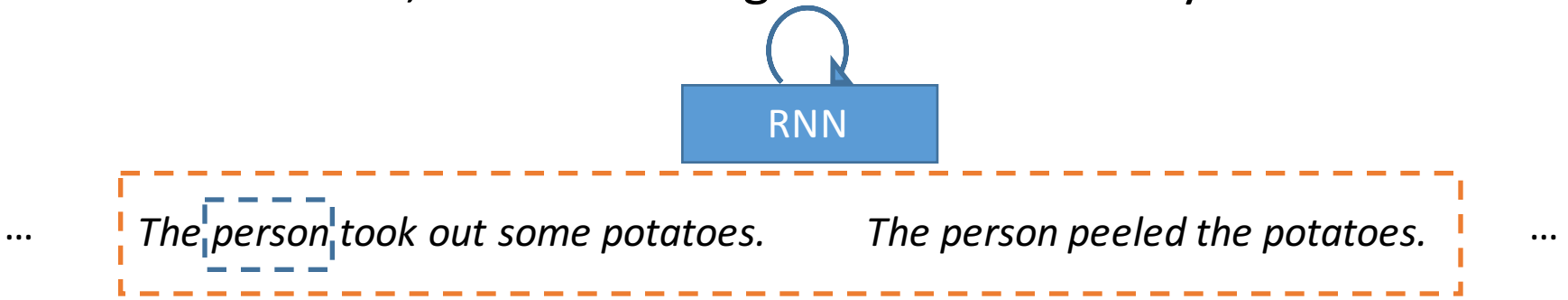
flat structure, concatenating sentences directly with one RNN



RNN-cat vs. h-RNN

RNN-cat

flat structure, concatenating sentences directly with one RNN



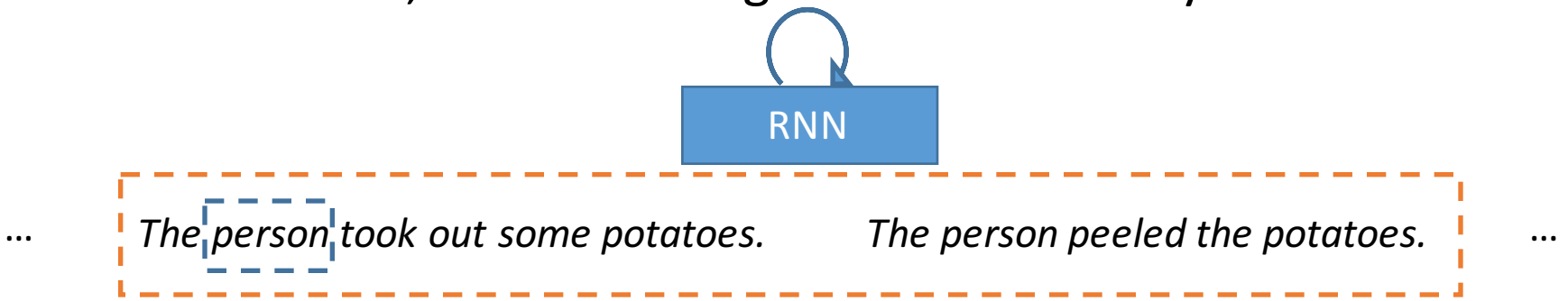
Amazon Mechanical Turk (AMT): side-by-side comparison

```
Which of the two sentences better describes the video?
 1. the first    2. the second. 3. Equally good or bad
```

RNN-cat vs. h-RNN

RNN-cat

flat structure, concatenating sentences directly with one RNN

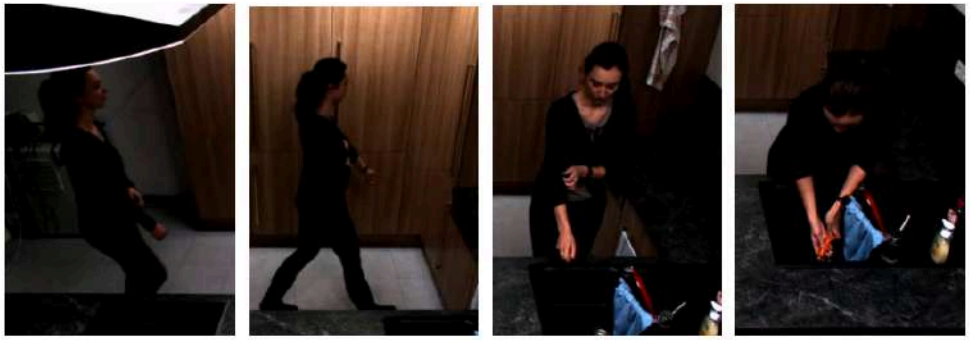


Amazon Mechanical Turk (AMT): side-by-side comparison

```
Which of the two sentences better describes the video?
 1. the first    2. the second. 3. Equally good or bad
```

h-RNN	RNN-cat	Equally good or bad	Total
773	472	3069	4314

RNN-sent vs. h-RNN examples



RNN-sent:	*The person entered the kitchen.* *The person went to the refrigerator.* *The person placed the cucumber on the cutting board.* *The person rinsed the cutting board.*
h-RNN:	*The person walked into the kitchen.* *The person went to the refrigerator.* *The person walked over to the sink.* *The person rinsed the carrot in the sink.*

RNN-sent:	*The person took out a cutting board from the drawer.* *The person got a knife and a cutting board from the drawer.* *The person cut the ends off the cutting board.*
h-RNN:	*The person took out a cutting board.* *The person got a knife from the drawer.* *The person cut the cucumber on the cutting board.*

Conclusions & Discussions

Hierarchical RNN improves paragraph generation

Conclusions & Discussions

Hierarchical RNN improves paragraph generation

Issues:

1. Most errors occur when generating nouns; small objects hard to recognize (on TACoS-MultiLevel)

2. One-way information flow

3. Language model helps, but sometimes overrides computer vision result in a wrong way

Thanks!

Baidu Research

Video Paragraph Captioning using Hierarchical Recurrent Neural Networks

Haonan Yu[1] Jiang Wang[1] Zhiheng Huang[2] Yi Yang[1] Wei Xu[1]
[1] Baidu Research – Institute of Deep Learning [2] Tencent

CVPR 2016

Problem

A dog is playing in a bowl.

A man is pouring oil into a pot.

Figure 1. Only one sentence is generated for a video with few details.

Why generating a paragraph?

Using only one short sentence to describe a semantically rich video usually yields uninformative and even boring results. For example, instead of saying "*the person sliced the potatoes, cut the onions into pieces, put the onions and potatoes into the pot, and turned on the stove*", a method that is only able to produce one short sentence would probably say "*the person is preparing food*".

The idea

We want to explicitly model the temporal dependency among sentences for multi-sentence generation. The generation of one sentence is affected by the semantic context given by previous sentences. For example, in a video of cooking dishes, a sentence "*the person peeled the potatoes*" is more likely to occur, than "*the person turned on the stove*", after "*the person took out some potatoes*".

Approach

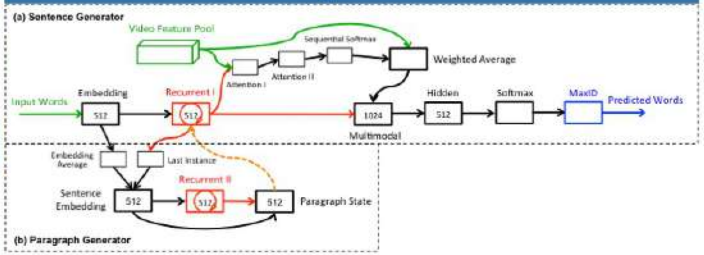


Figure 2. Our hierarchical RNN for video captioning. Green denotes the input to the framework, blue denotes the output, and red denotes the recurrent components. The orange arrow represents the reinitialization of the sentence generator with the current paragraph state.

Our approach stacks a paragraph generator on top of a sentence generator. The sentence generator is built upon:

1) a Recurrent Neural Network (RNN) for language modeling,
2) a multimodal layer for integrating information from different sources, and
3) an attention model for selectively focusing on the input video features.

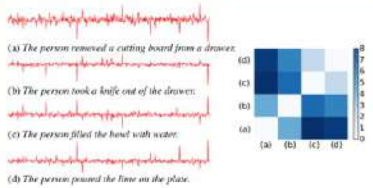


Figure 3. Left column: four examples of our learned sentence embeddings. Each 512 dimensional embedding is drawn as a red curve. Right column: the Euclidean distance between every two embeddings. A small distance indicates that two sentences have similar meanings. Notice how (a) and (b) are similar to each other due to sharing common keywords. Also note that even though (c) and (d) are quite different literally, our framework learns similar semantics for them from the video features.

Experiment results

The person opened the drawer.
The person took out a pot.
The person went to the sink.
The person washed the pot.
The person turned on the stove.

The person peeled the fruit.
The person put the fruit in the bowl.
The person sliced the orange.
The person put the pieces in the plate.
The person rinsed the plate in the sink.

Figure 4. Examples of our generated sentences. The video frames are cropped around the person only for better visualization.

We evaluate on two benchmark datasets: YouTubeClips [1] and TACoS-MultiLevel [2]. The YouTubeClips dataset contains 1,967 short videos with 80,839 sentences in total. The TACoS-MultiLevel dataset contains 185 long videos with 52,478 sentences in total. We employ three different evaluation metrics: BLEU, METEOR, and CIDEr.

	YouTubeClips		TACoS-MultiLevel		
	LSTM-E [3]	Our Method	LRCN [4]	RNN-cat	Our Method
BLEU@4	0.453	0.604	0.292	0.297	0.305
METEOR	0.310	0.326	0.282	0.284	0.287
CIDEr	N/A	0.658	1.534	1.555	1.602

References

1. D. L. Chen and W. B. Dolan. *Collecting highly parallel data for paraphrase evaluation*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 2011.
2. A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. *Coherent multi-sentence video description with variable level of detail*. In German Conference on Pattern Recognition (GCPR), 2014.
3. Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. *Jointly modeling embedding and translation to bridge video and language*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
4. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. *Long-term recurrent convolutional networks for visual recognition and description*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

Poster #4