Multi-Task Video Captioning with Video and Entailment Generation

Ramakanth Pasunuru & Mohit Bansal

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Video Captioning Task

Ground truth: A woman is slicing a red pepper.

Ground truth: A group of boys are fighting.

- Assistance to visually impaired
- Improving online video search
- Grounded robotic instruction tasks

[Kojima et al., 2002; Lee et al., 2008; Khan and Gotoh, 2012; Barbu et al., 2012; Das et al., 2013; Rohrbach et al., 2013; Yu and Siskind, 2013; Venugopalan et al., 2014, 2015, 2016]

Video Captioning Task

Ground truth: A woman is slicing a red pepper.

SotA Baseline: A woman is slicing a carrot.

Ground truth: A group of boys are fighting.

SotA Baseline: A group of men are dancing.

[Kojima et al., 2002; Lee et al., 2008; Khan and Gotoh, 2012; Barbu et al., 2012; Das et al., 2013; Rohrbach et al., 2013; Yu and Siskind, 2013; Venugopalan et al., 2014, 2015, 2016]

Video Captioning Task

Ground truth: A woman is slicing a red pepper.

SotA Baseline: A woman is slicing a carrot.

Our model: A woman is slicing a pepper.

Ground truth: A group of boys are fighting.

SotA Baseline: A group of men are dancing.

Our model: Two men are fighting.

[Kojima et al., 2002; Lee et al., 2008; Khan and Gotoh, 2012; Barbu et al., 2012; Das et al., 2013; Rohrbach et al., 2013; Yu and Siskind, 2013; Venugopalan et al., 2014, 2015, 2016]

Multi-Task Learning

- Paradigm to improve generalization performance of a task using related tasks.
- The multiple tasks are learned in parallel (alternating optimization mini-batches) while using shared model representations/parameters.
- Each task benefits from extra information in the training signals of related tasks.
- Luong et al., 2016 presented multi-task learning for sequence-to-sequence models, with shared encoder or decoder representations.

[Caruana, 1998; Argyriou et al., 2007; Kumar and Daume, 2012; Luong et al., 2016]

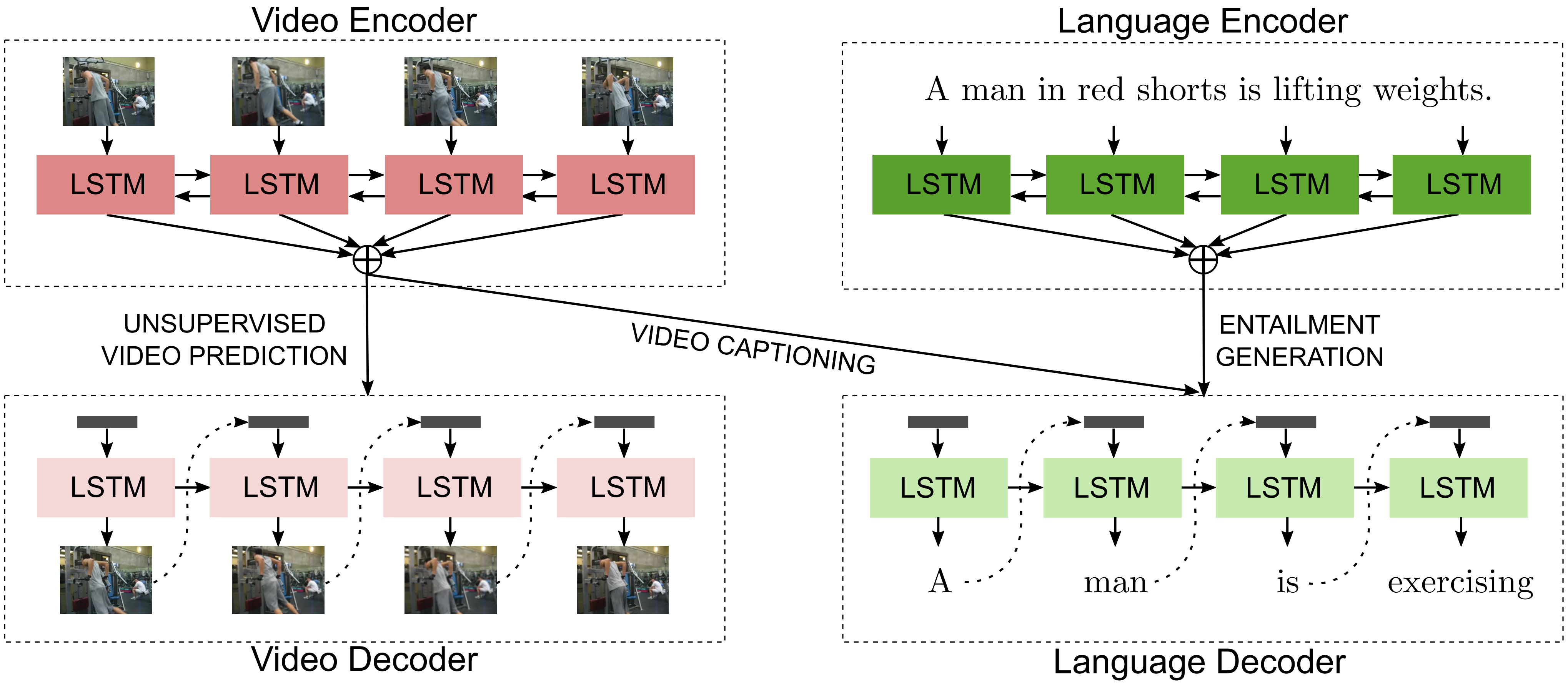
Multi-Task for Video Captioning

- Video Captioning Challenges:
  - Lack of sufficient labeled data
  - Spatial-visual modeling
  - <u>Logical</u> storyline dynamics
  - <u>Temporal</u> across-frame dynamics

Ground truth: A person is mixing powdered ingradients with water.
A woman is mixing flour and water in a bowl.
Our model: A woman is mixing ingredients in a bowl.

- We share knowledge w/ 2 related directed-generation tasks (textual+visual):
  1. Premise-to-Entailment Generation
     (to help learn better caption decoder representations, since caption is also entailed by video)
  2. Video-to-Video Generation (Unsupervised)
     (to help learn richer video encoder representations, aware of temporal action context)

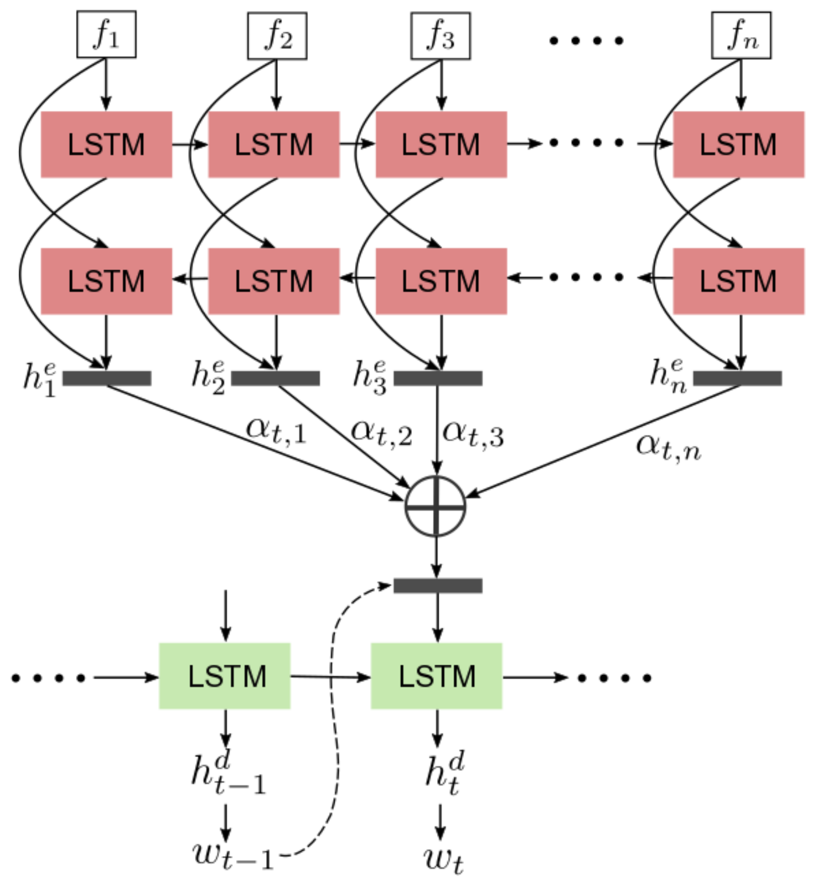
M-to-M Multi-Task for Video Captioning



- Training in alternate mini-batches: mixing ratio = $\frac{\alpha_v}{(\alpha_v+\alpha_f+\alpha_e)} : \frac{\alpha_f}{(\alpha_v+\alpha_f+\alpha_e)} : \frac{\alpha_e}{(\alpha_v+\alpha_f+\alpha_e)}$

Baseline Video Captioning Model

- Sequence-to-sequence encoder-decoder model
- Attention-based (Bahdanau et al., 2015)
- State-of-the-art Inception-v4 image frame features
- Strong baseline (>= previous work)



Textual Entailment

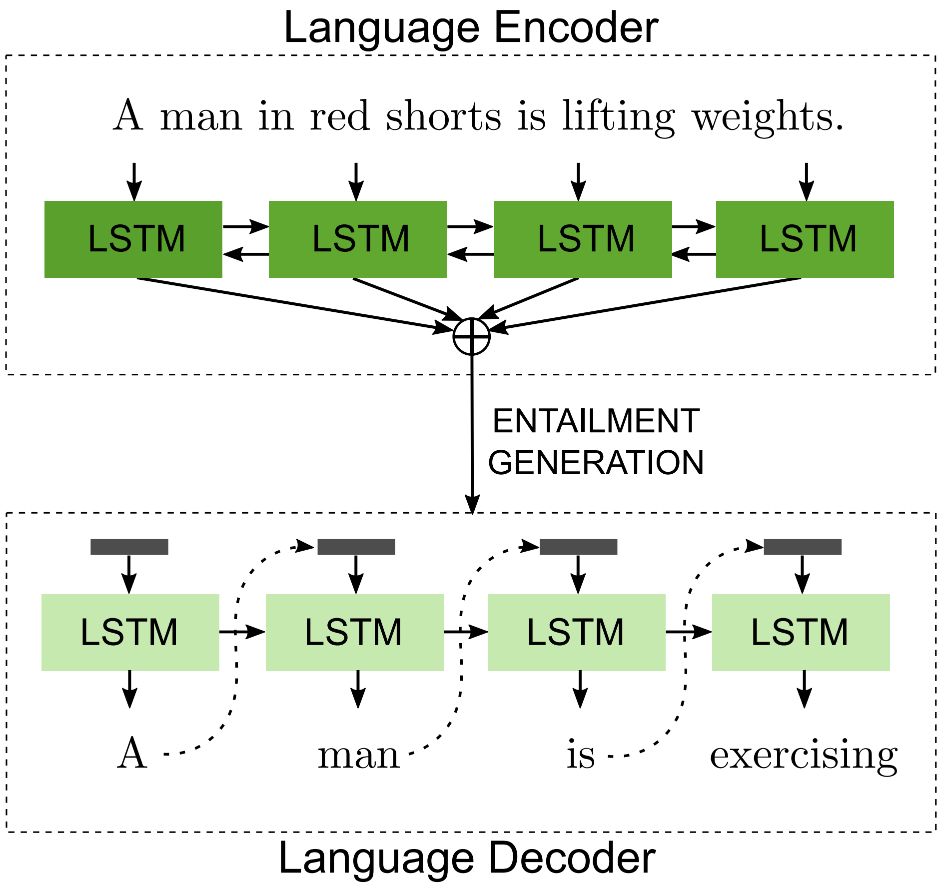
- Directional, logical-implication relation between two sentences:
  - Premise: *A girl is jumping on skateboard in the middle of a red bridge.*
  - Entailment: *The girl does a skateboarding trick.*
  - Contradiction: *The girl skates down the sidewalk.*
  - Neutral: *The girl is wearing safety equipment.*

  - Premise: *A blond woman is drinking from a public fountain.*
  - Entailment: *The woman is drinking water.*
  - Contradiction: *The woman is drinking coffee.*
  - Neutral: *The woman is very thirsty.*
- Can we use entailment as linguistic inference to help related directed/conditioned generation tasks? (Yes, for e.g. video captioning or document summarization)
- Large-scale SNLI corpus allows training accurate classification and RNN-style generation models

[Dagan at al., 2006; Roth and Sammons, 2007; Lai and Hockenmaier, 2014; Bowman et al., 2016]

Entailment Generation Model

- Helps learn better video-entailing caption decoder representations
- Since caption needs to be entailed by visual premise of video (i.e., describes subsets of objects/events logically implied by full video content), we teach it about entailment via MTL.
- Better than simply fusing an external LM to decoder (premise-to-entailment task matches logically-directed video-to-caption task better).



[Kolesnyk et al., 2016]

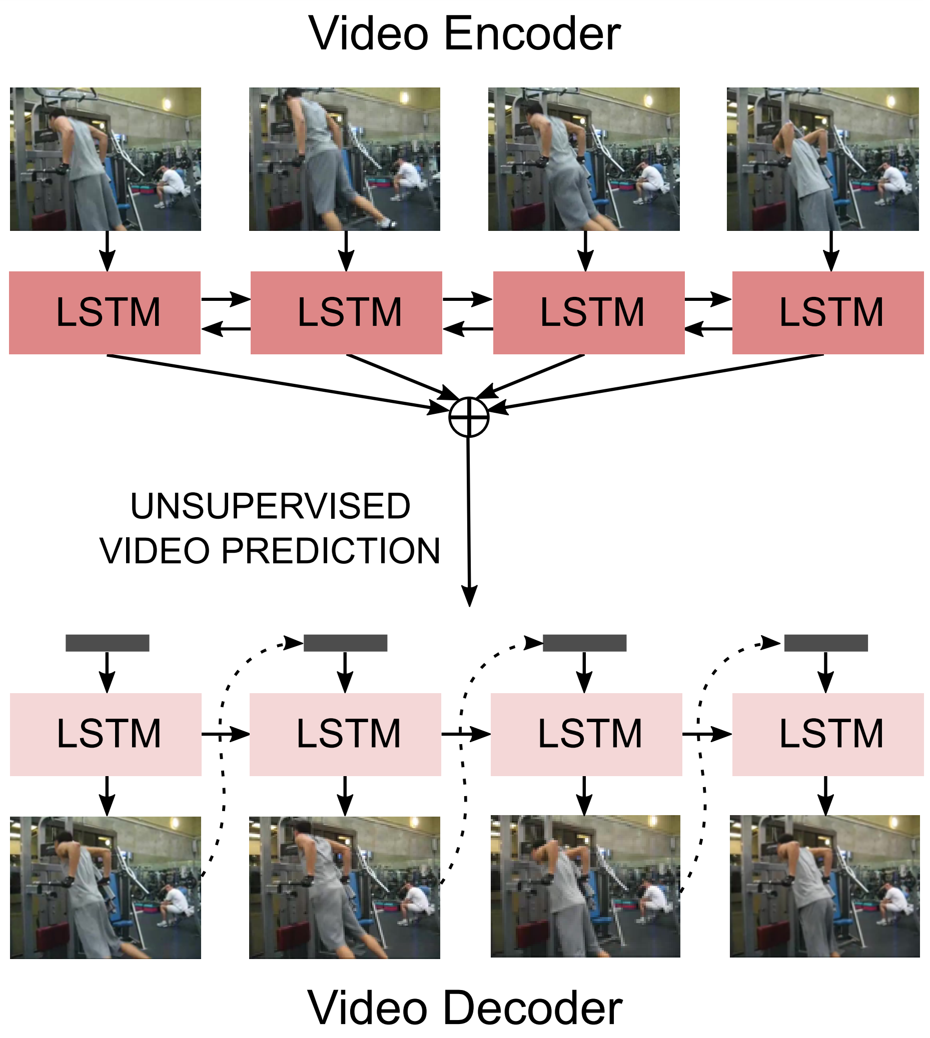
Unsupervised Video Prediction

- Helps learn richer video encoder representations that are aware of temporal context and action sequence/completion

- Robust to missing frames and varying frame lengths or motion speeds

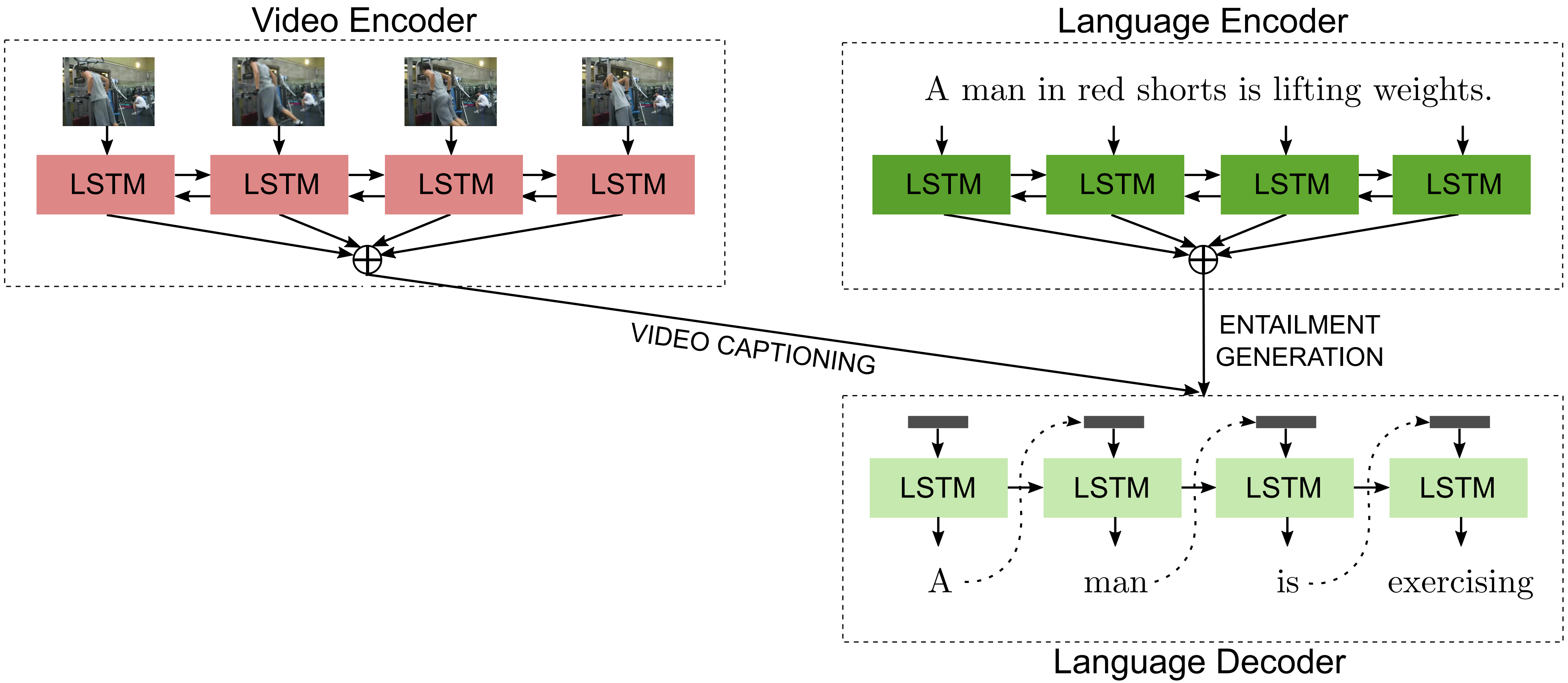
- 80:20% frame division between encoder and decoder

- UCF-101 action videos dataset

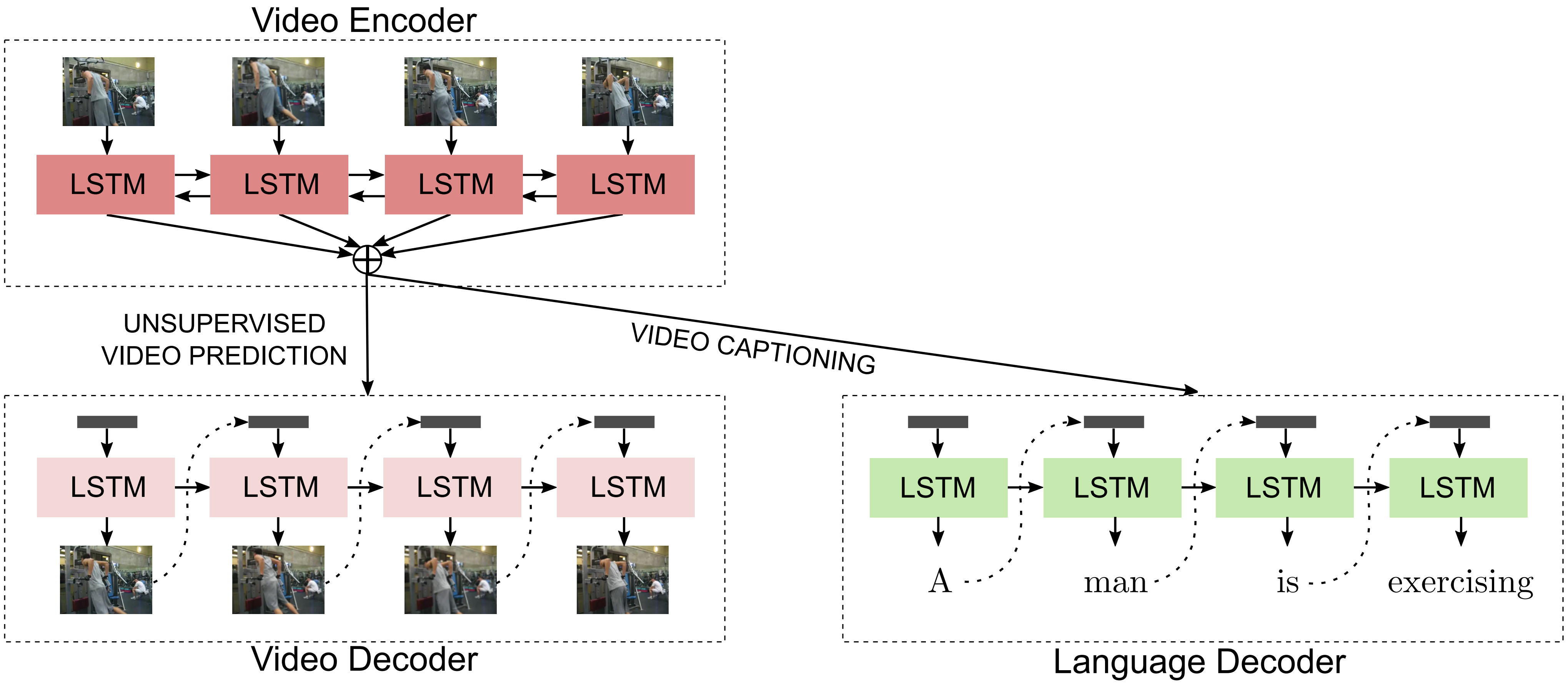


[Srivastava et al., 2015]

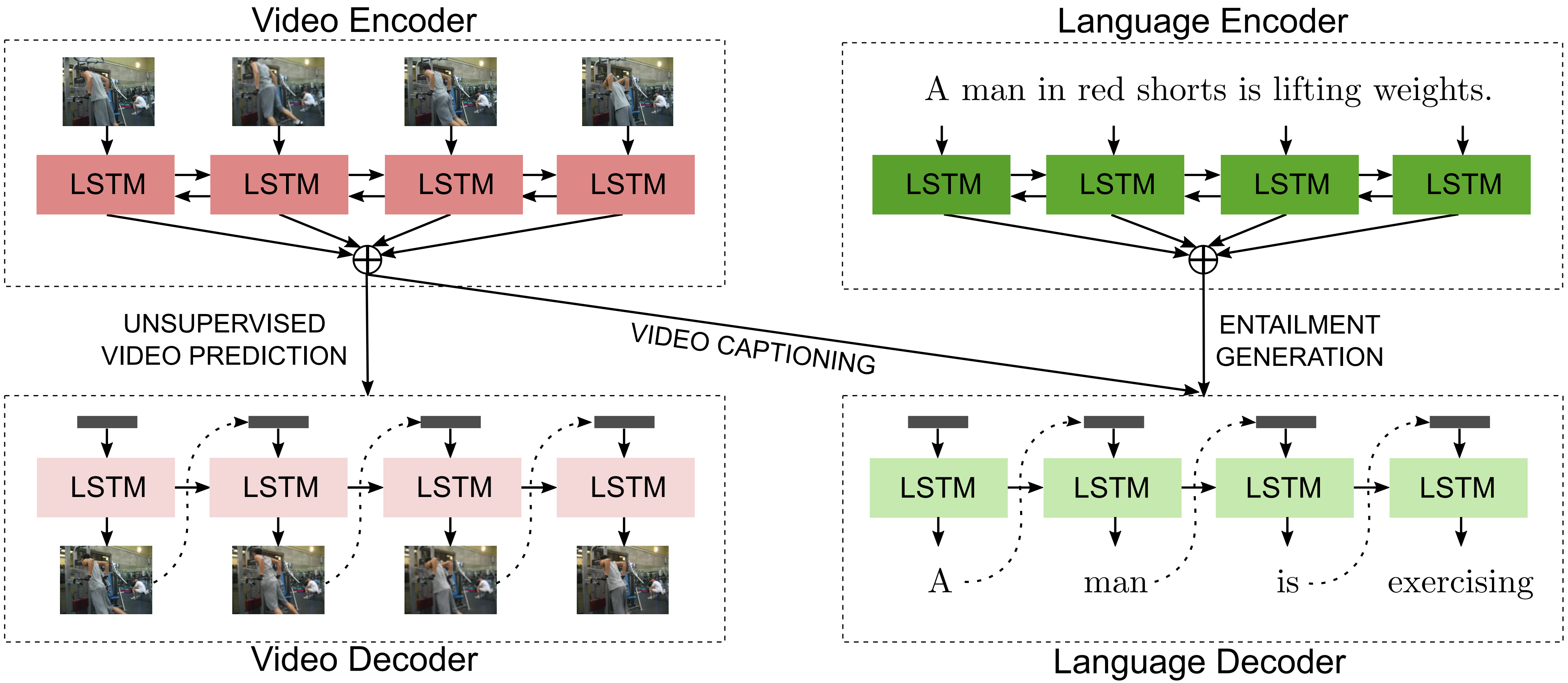
M-to-1 Multi-Task Model



1-to-M Multi-Task Model



M-to-M Multi-Task Model



- Training in alternate mini-batches: mixing ratio = $\frac{\alpha_v}{(\alpha_v+\alpha_f+\alpha_e)} : \frac{\alpha_f}{(\alpha_v+\alpha_f+\alpha_e)} : \frac{\alpha_e}{(\alpha_v+\alpha_f+\alpha_e)}$

Results (YouTube2Text/MSVD)

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
PREVIOUS WORK				
LSTM-YT (Venugopalan et al., 2015b)	26.9	-	-	31.2
S2VT (Venugopalan et al., 2015a)	29.8	-	-	-
Temporal Attention (Yao et al., 2015)	29.6	51.7	-	41.9
LSTM-E (Pan et al., 2016b)	31.0	-	-	45.3
Glove + DeepFusion (Venugopalan et al., 2016)	31.4	-	-	42.1
p-RNN (Yu et al., 2016)	32.6	65.8	-	49.9
HNRE + Attention (Pan et al., 2016a)	33.9	-	-	46.7

Results (YouTube2Text)

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
PREVIOUS WORK				
LSTM-YT (Venugopalan et al., 2015b)	26.9	-	-	31.2
S2VT (Venugopalan et al., 2015a)	29.8	-	-	-
Temporal Attention (Yao et al., 2015)	29.6	51.7	-	41.9
LSTM-E (Pan et al., 2016b)	31.0	-	-	45.3
Glove + DeepFusion (Venugopalan et al., 2016)	31.4	-	-	42.1
p-RNN (Yu et al., 2016)	32.6	65.8	-	49.9
HNRE + Attention (Pan et al., 2016a)	33.9	-	-	46.7
OUR BASELINES				
Baseline (V)	31.4	63.9	68.0	43.6
Baseline (G)	31.7	64.8	68.6	44.1
Baseline (I)	33.3	75.6	69.7	46.3
Baseline + Attention (V)	32.6	72.2	69.0	47.5
Baseline + Attention (G)	33.0	69.4	68.3	44.9
Baseline + Attention (I)	33.8	77.2	70.3	49.9
Baseline + Attention (I) (E) $\otimes$	35.0	84.4	71.5	52.6

Results (YouTube2Text)

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
PREVIOUS WORK				
LSTM-YT (Venugopalan et al., 2015b)	26.9	-	-	31.2
S2VT (Venugopalan et al., 2015a)	29.8	-	-	-
Temporal Attention (Yao et al., 2015)	29.6	51.7	-	41.9
LSTM-E (Pan et al., 2016b)	31.0	-	-	45.3
Glove + DeepFusion (Venugopalan et al., 2016)	31.4	-	-	42.1
p-RNN (Yu et al., 2016)	32.6	65.8	-	49.9
HNRE + Attention (Pan et al., 2016a)	33.9	-	-	46.7
OUR BASELINES				
Baseline (V)	31.4	63.9	68.0	43.6
Baseline (G)	31.7	64.8	68.6	44.1
Baseline (I)	33.3	75.6	69.7	46.3
Baseline + Attention (V)	32.6	72.2	69.0	47.5
Baseline + Attention (G)	33.0	69.4	68.3	44.9
Baseline + Attention (I)	33.8	77.2	70.3	49.9
Baseline + Attention (I) (E) ⊗	35.0	84.4	71.5	52.6
OUR MULTI-TASK LEARNING MODELS				
⊗ + Video Prediction (1-to-M)	35.6	88.1	72.9	54.1
⊗ + Entailment Generation (M-to-1)	35.9	88.0	72.7	54.4
⊗ + Video Prediction + Entailment Gener (M-to-M)	**36.0**	**92.4**	**72.8**	**54.5**

* All models (1-to-M, M-to-1 and M-to-M) stat. signif. better than strong SotA baseline.

Results (MSR-VTT)

- Diverse video clips from a commercial video search engine

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
Venugopalan et al., 2015	23.4	-	-	32.3
Yao et al., 2015	25.2	-	-	35.2
Xu et al., 2016	25.9	-	-	36.6
Rank1: v2t_navigator	28.2	44.8	**60.9**	**40.8**
Rank2: Aalto	26.9	45.7	59.8	39.8
Rank3: VideoLAB	27.7	44.1	60.6	39.1
Our Model (**New Rank1**)	**28.8**	**47.1**	60.2	**40.8**

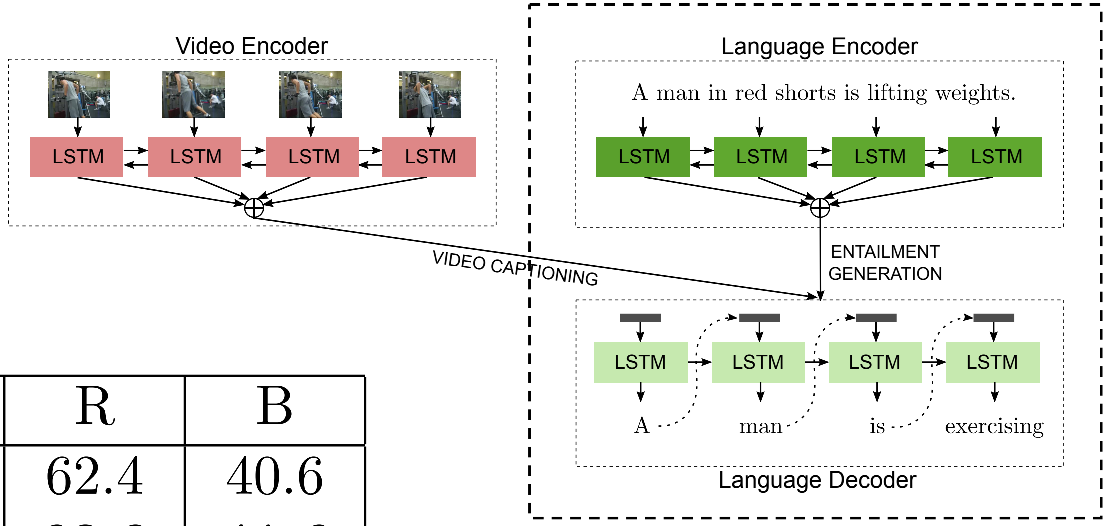
Results (MVAD)

- Movie video clips (1-2 human references so only METEOR feasible)

Models	METEOR
Yao et al., 2015	5.7
Venugopalan et al., 2015	6.7
Pan et al., 2016	6.8
Our M-to-M Multi-Task Model	**7.4**

Results (Entailment Generation)

- Video captioning mutually also helps improve the entailment-generation task in turn (w/ statistical significance)



Models	M	C	R	B
Entailment Generation	29.6	117.8	62.4	40.6
+Video Caption (M-to-1)	**30.0**	**121.6**	**63.9**	**41.6**

- New multi-reference split setup of SNLI to allow automatic metric evaluation and a zero train-test premise overlap

Human Evaluation

- Pilot human evaluations on 300-sized samples

- Multi-task model > strong non- multitask baseline on relevance and coherence/fluency (for both video captioning and entailment generation)

	YouTube2Text	
	Relev.	Coher.
Not Distinguish.	70.7%	92.6%
SotA Baseline Wins	12.3%	1.7%
Multi-Task Wins	**17.0%**	**5.7%**

Human Evaluation

- Pilot human evaluations on 300-sized samples

- Multi-task model > strong non- multitask baseline on relevance and coherence/fluency (for both video captioning and entailment generation)

	YouTube2Text		Entailment	
	Relev.	Coher.	Relev.	Coher.
Not Distinguish.	70.7%	92.6%	84.6%	98.3%
SotA Baseline Wins	12.3%	1.7%	6.7%	0.7%
Multi-Task Wins	**17.0%**	**5.7%**	**8.7%**	**1.0%**

Analysis Examples

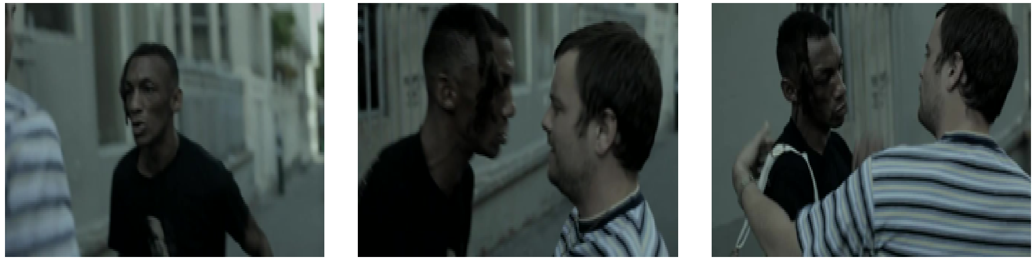
Ground truth: Two women are shopping in a store.
Two girls are shopping.
Baseline model: A man is doing a monkey in a store.
Multi-task model: A woman is shopping in a store.

Ground truth: Two men are fighting.
A group of boys are fighting.
Baseline model: A group of men are dancing.
Multi-task model: Two men are fighting.

(a) complex examples where the multi-task model performs better than baseline

Analysis Examples

Ground truth: A woman slices a shrimp tail.
A girl is cutting a fish tale.
Baseline model: A person is cutting the something.
Multi-task model: A woman is cutting a piece of meat.



Ground truth: Two men are talking aggressively.
The boy is talking.
Baseline model: A man is crying.
Multi-task model: A man is talking.

(b) ambiguous examples (i.e., ground truth itself confusing) where multi-task model still correctly predicts one of the possible categories

Analysis Examples

Ground truth: A monkey and a deer are fighting.
A gazelle is fighting with a baboon.
Baseline model: A man is walking on the ground.
Multi-task model: A monkey is walking.

Ground truth: A dog climbs into a dryer.
A dog is in a washing machine.
Baseline model: A man is playing.
Multi-task model: A man is playing with a toy.

(c) complex examples where both models perform poorly

(d) baseline > MTL: both correct but low specificity

- Overall, multi-task model's captions are better at both temporal action prediction and logical entailment w.r.t. ground truth captions (ablated examples in paper).

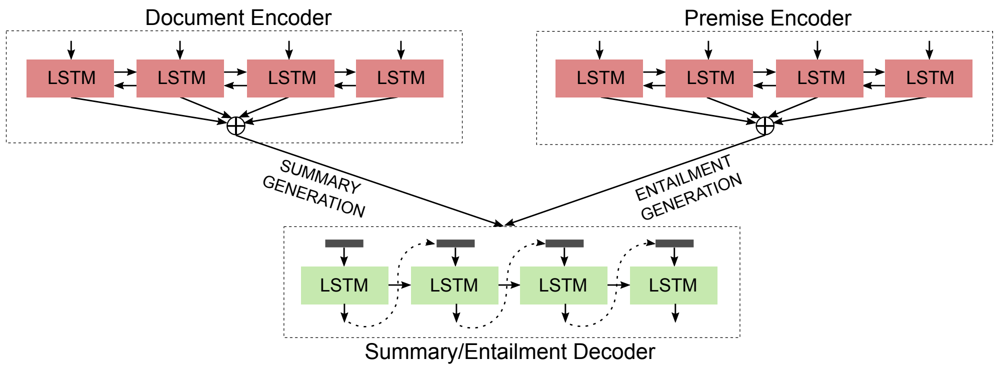
Entailment Generation Examples

Given Premise	Generated Entailment
a child that is dressed as spiderman is ringing the doorbell	a child is dressed as a superhero
a girl in cargo pants and a green shirt jumps in front of a square stone	a girl is jumping
a man in a red jacket rides a horse in mountainous terrain	a man is riding a horse
a woman in a dress with two children	a woman is wearing a dress
woman in a red headscarf covering her face	a woman is wearing a red scarf

Extensions and New Work

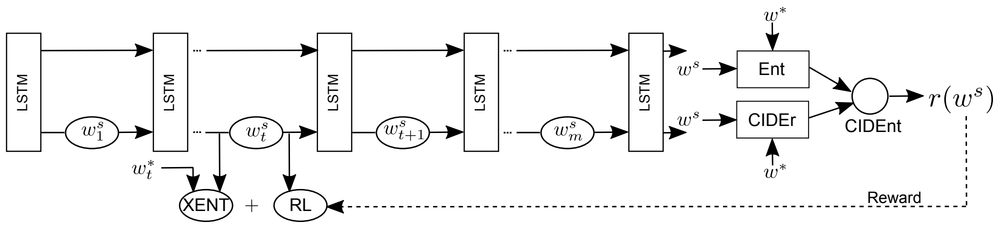
- Multitask Summarization with Entailment [EMNLP'17 – NewSumm]

  (A summary of a document is entailed by it)



- Entailment as reward in RL [EMNLP'17]

  (Corrects matching-based metrics to ensure logically-directed match and avoid contradiction)



Thanks!