Appendices for
Long-term Recurrent Convolutional Networks for Visual Recognition and Description

This supplemental material provides additional examples and more detailed description and analysis of our approach to activity recognition (Section A), image description (Section B) and video description (Section C). For the latter one we also provide results in the form of videos with subcaptions in the subfolder `video/`.

A. Activity Recognition

We report more detailed results on activity recognition on the UCF-101 dataset. First, we analyze which classes the LRCN improves the most on for split-1. We then report results on all three splits on the UCF-101 dataset and compare our system to other deep activity recognition models.

A.1. Accuracy for individual classes

Table 6 records the difference in accuracies between the LRCN model and single frame model for individual classes. We report the 10 classes in which the LRCN model improves most, and the 10 classes in which the LRCN model improves least. Recall that in the single frame model a CNN is fine-tuned with individual video frames. Video classification is done by averaging the predictions over all frames in a video. In the LRCN model, an LSTM receives inputs from the first fully connected layer of a CNN. The LRCN predicts an output for every frame, and final classification is also done by averaging over all the outputs of the LRCN.

For the majority of classes LRCN improves performance over the single frame model. Though the LRCN performs worse on some classes including knitting and mixing, in general when the LRCN performs worse, the loss in accuracy is not as substantial as the gain in accuracy in classes such as “jump rope” and “push ups”. Thus, accuracy is higher overall.

Table 7 records the difference in accuracies between classes for the LRCN flow and LRCN RGB model. This demonstrates that RGB and flow are helpful in classifying different types of actions and thus helps explain the boost in performance seen when averaging RGB and flow models.

Figure 5 shows how learning the dynamics between frames impacts outputs for the video `v_JumpRope_g06_c01.avi` when flow is used as an input. Note that many of the flow frames (3, 5, 6, and 9) have near zero flow and thus appear a solid gray. Figure 5 right shows the LRCN model outputs at each time step of the LRCN and figure 5 left shows the output of the single frame model. The prediction for the first frame for each model is the same (*WriteOnBoard*). However, by the fifth frame, the LRCN model recognizes the video as *JumpRope* and all other subsequent outputs of the LRCN are labeled *JumpRope*. The single frame model is unable to remember previous frames resulting in less consistent predictions. After averaging over all frames, the LRCN predicts the correct label, *JumpRope*.

A.2. Evaluation for UCF-101 over all splits

We report results on UCF-101 splits 1, 2 and 3 and its average on the single frame model and LRCN-fc$_6$ model (Table 8). The LRCN model helps substantially when the input to the system is flow frames, though does not consistently help when the input is RGB frames. The LRCN fusion model improves upon the single frame fusion model by approximately 4%. Recall that to fuse the RGB and flow inputs, we train a network on each input type. We then compute a weighted average (0.33 RGB/0.67 flow) for the RGB and flow networks as was done in [33]. When comparing our model to other deep activity recognition models, we note that we outperform [16] by over 15%, where authors could not improve much over their single-frame model. While we slightly underperform the state-of-the-art in CNN-based activity recognition [33], we expect that when visual activity recognition datasets evolve to require more sophisticated, long-term forms of reasoning, methods like LRCN that account for long-term dynamics will be advantageous.

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

Activity	Δ	Activity	Δ
JumpRope	34.21	Knitting	-20.59
PushUps	30.00	Mixing	-20.00
HighJump	29.73	PoleVault	-12.50
FieldHockeyPenalty	22.50	GolfSwing	-10.26
BoxingPunchingBag	22.45	WalkingWithDog	-8.33
YoYo	22.22	BrushingTeeth	-8.11
RopeClimbing	20.59	TableTennisShot	-7.69
HeadMassage	19.51	PlayingDaf	-7.31
ApplyLipstick	18.75	Skijet	-7.14
ApplyEyeMakeup	18.18	UnevenBars	-7.14

Table 6: Activity recognition: Comparing which classes do better and worse with LRCN model in comparison to single frame model. Here we report results on all three splits of UCF-101 (only results on the first split were presented in the paper).Δ is the difference between LRCN accuracy and frame accuracy.

Flow - RGB	+Δ	Flow - RGB	-Δ
SoccerJuggling	48.72	FieldHockeyPenalty	-57.50
BodyWeightSquats	46.67	TennisSwing	-51.02
PushUps	46.67	Typing	-44.19
Basketball	45.71	CuttingInKitchen	-36.36
JumpRope	44.74	BrushingTeeth	-36.11
BoxingPunchingBag	40.82	Skijet	-28.57
LongJump	38.46	Skiing	-27.50
HandstandWalking	38.23	FloorGymnastics	-25.00
ApplyEyeMakeup	36.36	BaseballPitch	-23.26
ShavingBeard	34.88	Mixing	-22.22

Table 7: Activity recognition: Comparing which classes do better and worse with LRCN RGB and LRCN flow models. Δ is the difference between LRCN flow accuracy and LRCN RGB accuracy.

	R@1	**R@5**	**R@10**	**Med*r***
LRCN$_{1u}$	14.1	31.3	39.7	24
LRCN$_{2u}$	3.8	12.0	17.9	80
LRCN$_{2f}$	**17.5**	**40.3**	**50.8**	**9**
LRCN$_{4f}$	15.8	37.1	49.5	10

Table 9: Flickr30k caption-to-image retrieval results for variants of the LRCN architectures. See Figure 6 for diagrams of these architectures. The results indicate that the "factorization" is important to the LRCN's retrieval performance, while simply stacking additional LSTM layers does not seem to improve performance.

B. Image Description

B.1. Architectural ablation

In Table 9, we report image-to-caption retrieval results for each of the architectural variants in Figure 6, as well as a four-layer version (LRCN$_{4f}$) of the factored model. Based on the facts that LRCN$_{2f}$ outperforms the LRCN$_{4f}$ model, and LRCN$_{1u}$ outperforms LRCN$_{2u}$, there seems to be little gained by naively stacking additional LSTM layers atop an existing network. On the other hand, a comparison of the LRCN$_{2f}$ and LRCN$_{2u}$ results indicates that the "factorization" in the architecture is quite important to the model's retrieval performance.

B.2. Sample captions

We display the first 24 images from our randomly selected COCO [24] validation subset and the corresponding captions generated by our fine-tuned LRCN model.

The generated sentences are usually grammatically correct and, with few exceptions among these 24, accurately describe the content of the image. Subjectively, one of the most common and striking sources of error in the captions is the misidentification of human characteristics like gender ("his" instead of "her") and age ("woman" instead of "girl").

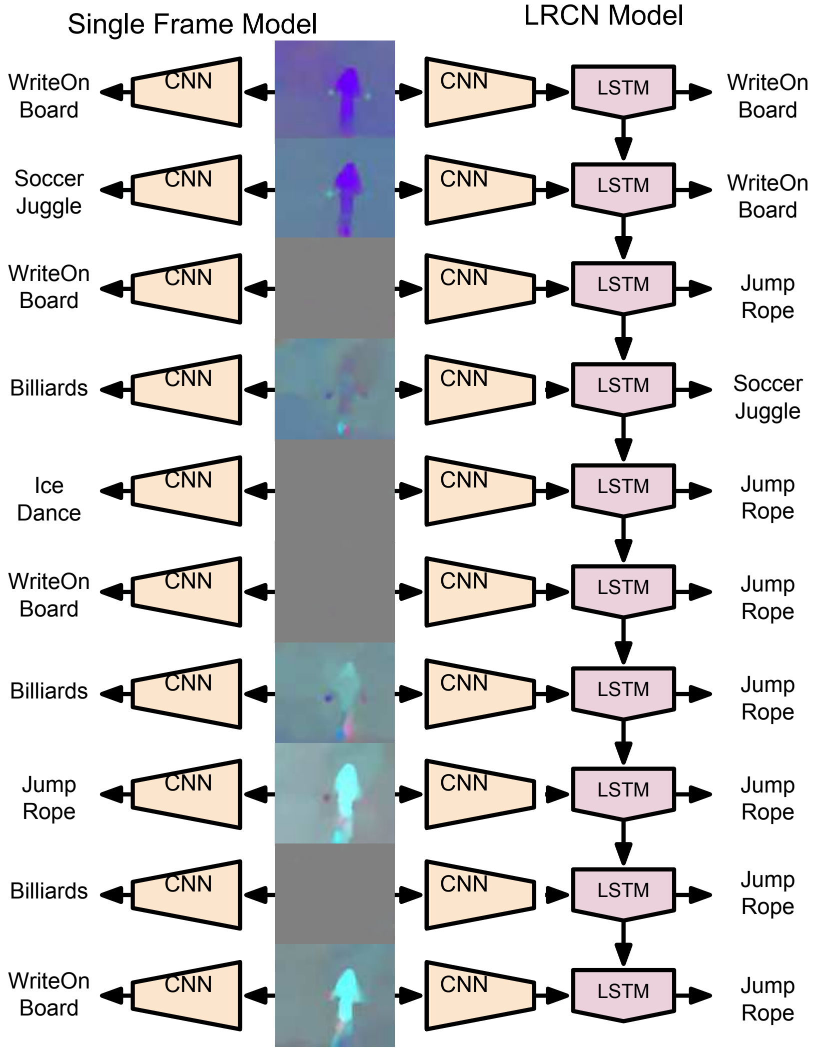


Figure 5: Activity recognition: Comparison of outputs of the LRCN video (right) and single frame model (left) for the video `v_JumpRope_g06_c01.avi`. The LRCN learns the correct label by frame 5 of the video and outputs this label for the remaining frames. Without information from previous hidden states (left), the labels are not as accurate or stable.

Model	**RGB**				**Flow**				**Fusion**			
	Spl. 1	Spl. 2	Spl. 3	Avg.	Spl. 1	Spl. 2	Spl. 3	Avg.	Spl. 1	Spl. 2	Spl. 3	Avg.
Single frame	65.40	64.62	63.08	64.30	53.20	55.46	54.35	54.34	–	–	–	–
Single frame (average)	69.02	**66.79**	**67.29**	67.70	72.19	72.44	71.94	72.19	79.04	79.03	78.46	78.84
LRCN-fc$_6$	**71.12**	66.71	66.75	**68.19**	**76.95**	**77.24**	**78.20**	**77.46**	**82.92**	**82.40**	**82.66**	**82.66**

Table 8: Activity recognition: Comparing single frame models to LRCN networks for activity recognition in the UCF-101 [37] dataset, with both RGB and flow inputs. Our LRCN model consistently outperforms a model based on predictions from the underlying convolutional network architecture alone.

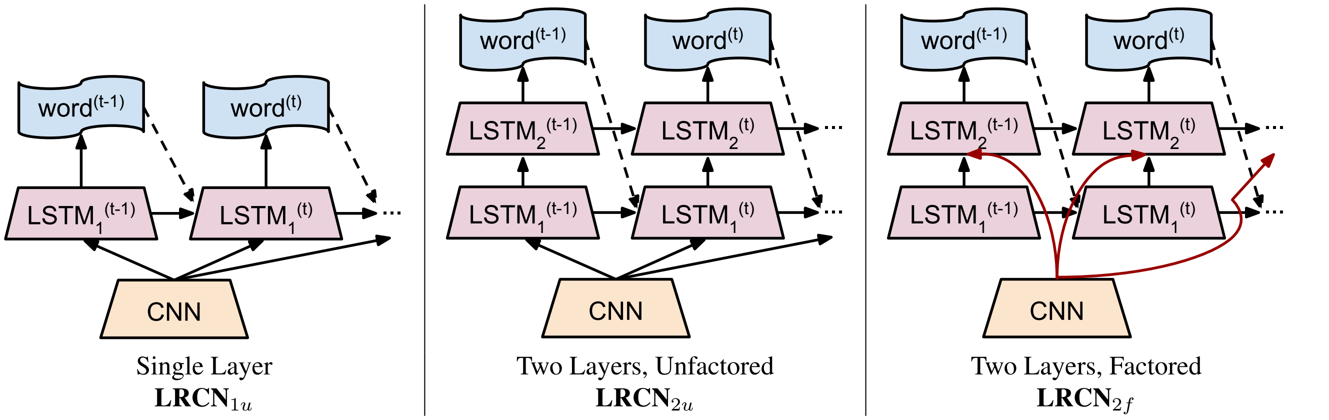
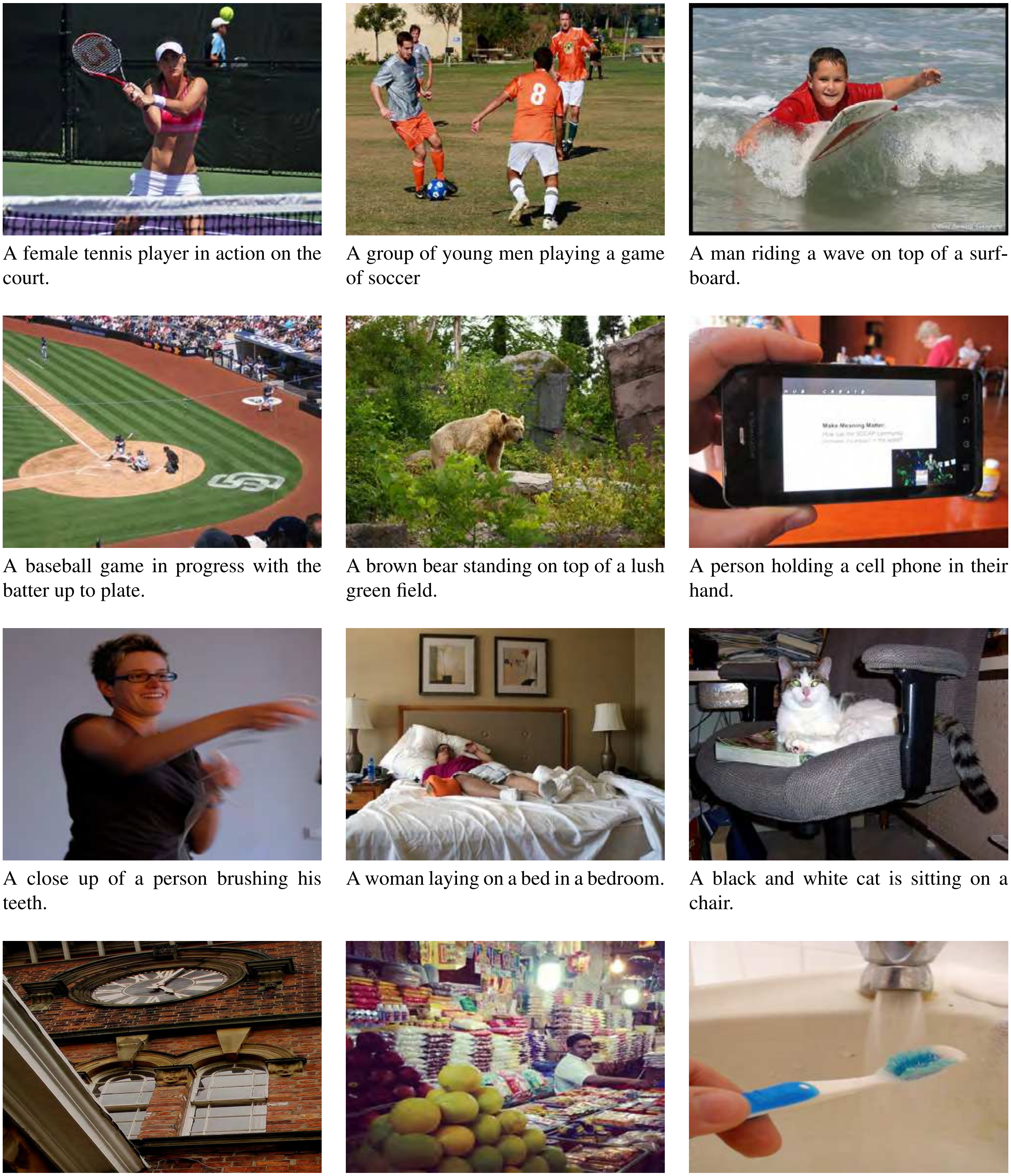


Figure 6: Three variations of the LRCN image captioning architecture that we experimentally evaluate. We explore the effect of depth in the LSTM stack, and the effect of the "factorization" of the modalities (also explored in [19]).



A female tennis player in action on the court.

A group of young men playing a game of soccer

A man riding a wave on top of a surfboard.

A baseball game in progress with the batter up to plate.

A brown bear standing on top of a lush green field.

A person holding a cell phone in their hand.

A close up of a person brushing his teeth.

A woman laying on a bed in a bedroom.

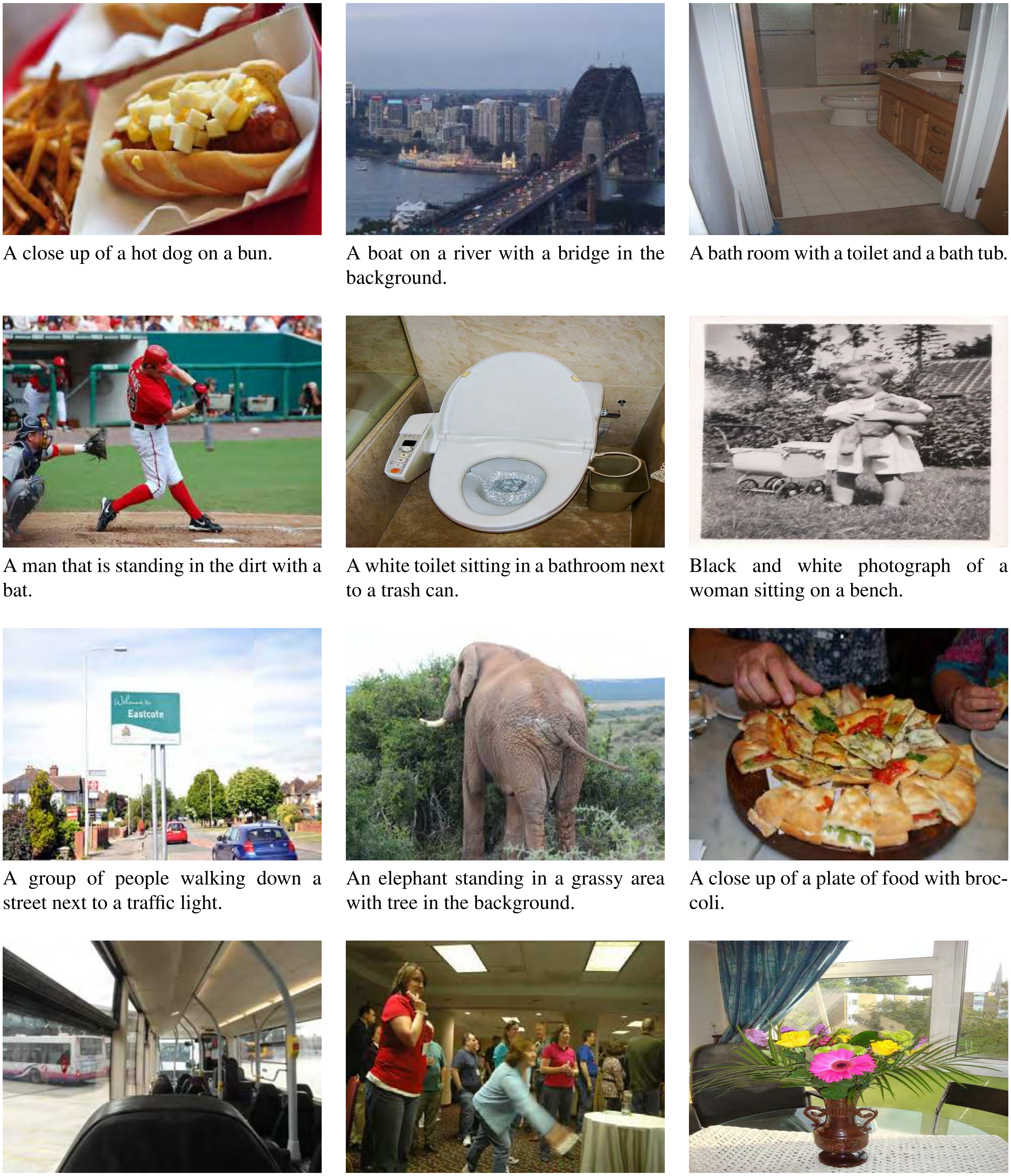
A black and white cat is sitting on a chair.

A large clock mounted to the side of a building.

A bunch of fruit that are sitting on a table.

A toothbrush holder sitting on top of a white sink.

Figure 7: Image description: images with corresponding captions generated by our finetuned LRCN model. These are images 1-12 of our randomly chosen validation set from COCO 2014 [24] (see Figure 8 for images 13-24). We used beam search with a beam size of 5 to generate the sentences, and display the top (highest likelihood) result above.



A close up of a hot dog on a bun.

A boat on a river with a bridge in the background.

A bath room with a toilet and a bath tub.

A man that is standing in the dirt with a bat.

A white toilet sitting in a bathroom next to a trash can.

Black and white photograph of a woman sitting on a bench.

A group of people walking down a street next to a traffic light.

An elephant standing in a grassy area with tree in the background.

A close up of a plate of food with broccoli.

A bus parked on the side of a street next to a building.

A group of people standing around a table.

A vase filled with flower sitting on a table.

Figure 8: Image description: images 13-24 (and LRCN-generated captions) from the set described in Figure 7.

C. Video Description

This section illustrates our video description in more detail and with more examples.

C.1. Extended approach figure

We provide a larger and slightly more detailed version of Figure 4 in the submission in Figure 10, comparing our different variants for video description. It tries to show more clearly the difference between variant (a) and (b). In (a), the input is given sequentially to the encoder LSTM (orange), handling different length input. We can have different length input as a CRF state might be encoded with two words (in the example cutting-board) or with zero words due to a null state. We do this to follow most closely the input and output given to the SMT Moses Toolkit in [29]. In contrast variant (b) uses a fixed length input, which is replicated to all time steps. And (c) differs to (b) by using the the probability distribution of the CRF rather than the max decoding output.

C.2. Supplemental video examples

To show case our video description we provide example videos. They can be found in the `videos/` subfolder. The descriptions are blended in as subtitles. As the playback speed is 10x of the original speed (to reduce space and playing time) some video segments described appear very shortly and in rare cases sentences overlap; please pause the video in case you would like to analyze the generated sentences more extensively. See the next section for a discussion.

We confirmed that our videos can be opened using the following platforms:

Mac VLC Player (http://www.videolan.org/vlc/, works also for Windows/Linux) or QuickTime Player (after an automatic conversion).

Windows Windows media player.

Linux mplayer.

We note that we use ground truth temporal segments from [29] in order to compare to human descriptions, but our approach also runs with automatic segmentation as in [29].

C.3. Video description: discussion of qualitative results

While we evaluate sentences per video clip we examine descriptions of a full video in this section. This shows that our approach is able to produce consistent multi-sentence descriptions for an entire video. We generate a sentence for each video clip, but we use the semantic representation, which forms the input of the LSTM, to model consistency across sentences. The consistency is modeled in a CRF [29] by using a common topic across all sentences and relying on full video level features. Modeling that the sentences are consistent with the overall topic (in the kitchen scenario the dish, *e.g.* "preparing scrambled egg") reduces the topic switches, *i.e.* we reduce the cases of saying *e.g.* "*The person cracked an egg.*" and in the next sentences "*Then, the person put the carrots in the pot.*" rather than "*Then, the person poured the eggs in the pan.*". As mentioned in the submission we rely on the output of [29] for the semantic representation.

We provide three videos from TACoS Multilevel [29], for two videos (`s33-d27-preparing-onions.avi`, `s34-d69-scrambled-egg.avi`) the sentences are mainly correct, while `s29-d71-making-hot-dog---partial-failure -case.avi` shows many wrong objects and ingredients. The main reason for this is the incorrect visual recognition of the dish/topic with the CRF.

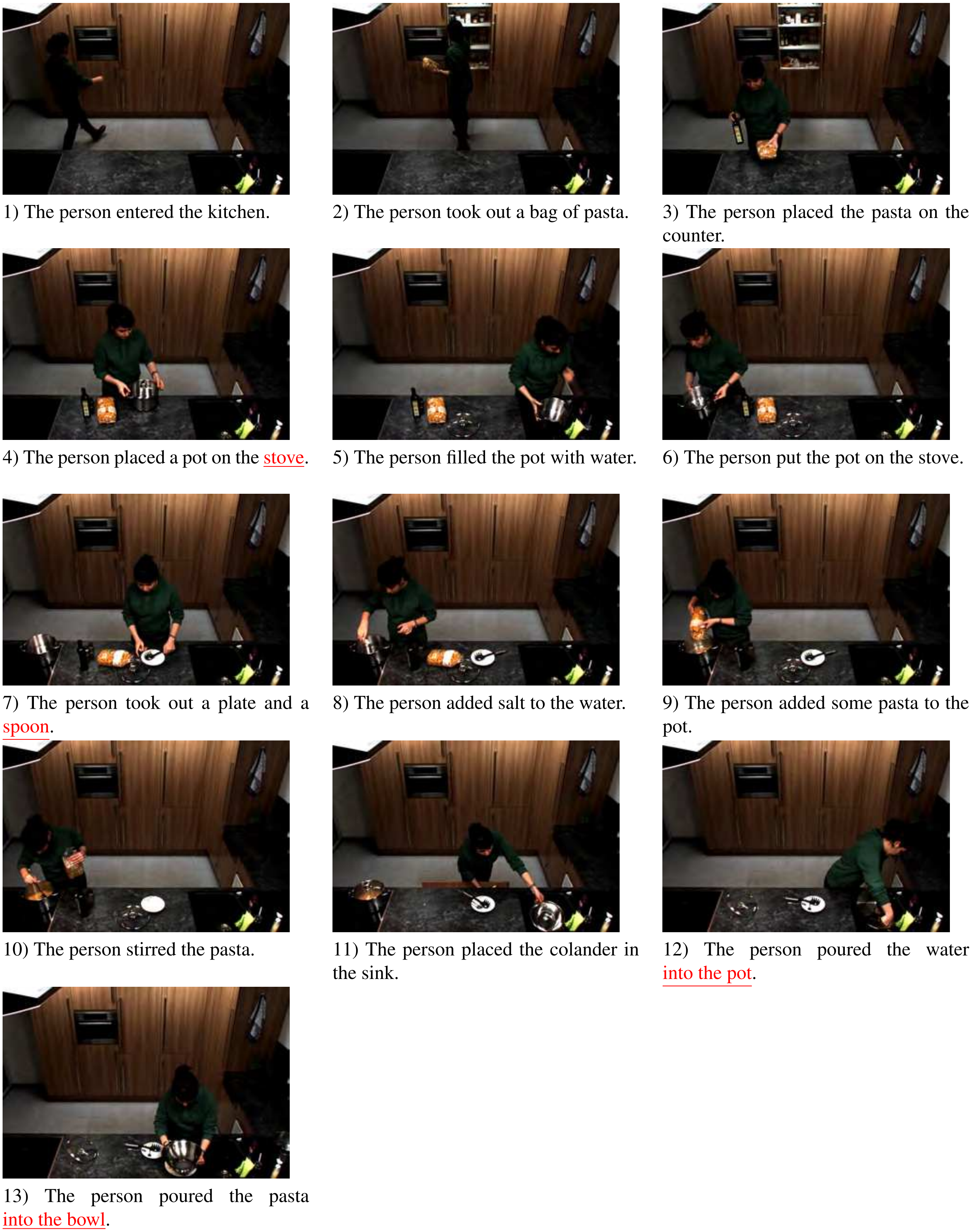
Figure 9 shows a sequence for *preparing pasta*. We show the output of our best performing variant (c). We notice that the description is overall correct, although it might miss things, *e.g.* the oil in sentence 3. Errors are made mainly for locations, *e.g. stove* versus *counter* in sentence 4, and fine grained object differences, *e.g. spoon* versus *pasta spoon* in sentence 7.

Table 10 compares the best version of our approach (c) with the other variants and [29] for a full cooking sequence for *preparing leeks*. Despite the same semantic representation, we notice several interesting aspects when comparing the different generation approaches and the human description.

1. While the human description is the most accurate description, it contains spelling errors, *e.g. retreaved* instead of *retrieved* or *chopped the root of* instead of *chopped the root off*. The translation systems typically do not make any spelling mistakes as these are unlikely according to the training data. However, they can make grammatical errors, in most cases by adding phrases or leaving words out.

2. While [29] switches the ingredient incorrectly between *cucumber, cauliflower, and egg*, variant (c) of our system consistently produces the correct ingredient.

3. Comparing our variants (a), (b) and (c), we notice for example, that for video snippet 6), 7), and 8), (a) and (b) produce the identical sentence *"The person cut the broccoli into small pieces."* / *"The person cut the leek in half."*, while (c) gives a more detailed description: *"The person cut the top off the leek."*, *"The person cut the top off the top"*. *"The person cut the top of the leek in half."*. While the variants (a) and (b) rely on a single max prediction, (c) can rely on the more detailed probability distribution and thus distinguish fine grained details.



1) The person entered the kitchen.

2) The person took out a bag of pasta.

3) The person placed the pasta on the counter.

4) The person placed a pot on the stove.

5) The person filled the pot with water.

6) The person put the pot on the stove.

7) The person took out a plate and a spoon.

8) The person added salt to the water.

9) The person added some pasta to the pot.

10) The person stirred the pasta.

11) The person placed the colander in the sink.

12) The person poured the water into the pot.

13) The person poured the pasta into the bowl.

Figure 9: Video description: Multi-sentence description of our approach (variant c) for a video showing *preparing pasta*. For subtitled video examples see `video/*.avi` and for a per sentence analysis see Table 10. Discussion in Section C. Red underlined marks errors.

1)		[29]	The person walked into the kitchen.
		ours (a)	The person entered the kitchen.
		ours (b)	The person entered the kitchen.
		ours (c)	The person entered the kitchen.
		human	The person entered the kitchen,
2)		[29]	The person took out a cutting board and knife from the drawer.
		ours (a)	The person took out a cutting board and a knife.
		ours (b)	The person took out a cutting board and a knife.
		ours (c)	The person took out a cutting board and a knife.
		human	and retreaves a cutting board and a knife.
3)		[29]	The person turned the broccoli in the sink.
		ours (a)	The person took out a bowl.
		ours (b)	The person put the leek on the cutting board.
		ours (c)	The person took out a leek.
		human	The person grabbed a leek from the fridge.
4)		[29]	The person cut the leek in half.
		ours (a)	The person put the chopped herbs into the bowl.
		ours (b)	The person chopped the leek into small pieces.
		ours (c)	The person cut the leek in half.
		human	The person proceed to remove the green leaves,
5)		[29]	The person washed the leek in the sink.
		ours (a)	The person washed the carrot.
		ours (b)	The person washed the leek in the sink.
		ours (c)	The person washed the leek.
		human	and rinsed the plant.
6)		[29]	The person cut the leek on the cutting board.
		ours (a)	The person cut the broccoli into small pieces.
		ours (b)	The person cut the leek in half.
		ours (c)	The person cut the top off the leek.
		human	The person chopped the root of
7)		[29]	The person cut the cucumber in half.
		ours (a)	The person cut the broccoli into small pieces.
		ours (b)	The person cut the leek in half.
		ours (c)	The person cut the top off the top.
		human	and sliced the leek down the middle.
8)		[29]	The person sliced the cauliflower on the cutting board.
		ours (a)	The person cut the broccoli into small pieces.
		ours (b)	The person cut the leek in half.
		ours (c)	The person cut the top of the leek in half.
		human	The person diced the leek into small sections.
9)		[29]	The person took out a frying pan from the drawer.
		ours (a)	The person took out a bowl.
		ours (b)	The person took out a leek.
		ours (c)	The person took out a plate and oil.
		human	The person retreaved a pan and butter,
10)		[29]	The person put the chopped onion on the cutting board on the cutting board.
		ours (a)	The person put the chopped herbs into the bowl.
		ours (b)	The person chopped the leek into small pieces.
		ours (c)	The person put the chopped leek into the pan.
		human	and added the leek to the hot pan.
11)		[29]	The person took out a spice from the spice rack,
		ours (a)	The person took a spice from the spice.
		ours (b)	The person took out a spice.
		ours (c)	The person took out two spices.
		human	The person retreaved spices,
12)		[29]	The person shook the to the frying pan.
		ours (a)	The person stirred the potatoes.
		ours (b)	The person stirred the leek.
		ours (c)	The person stirred the pan with a spatula.
		human	and stirred the leeks.
13)		[29]	The person took an egg from the cabinet.
		ours (a)	The person took out a broccoli.
		ours (b)	The person added some oil to the pan.
		ours (c)	The person added some oil to the pan.
		human	The person added water to the pan with broth and seasoning.
14)		[29]	The person melted butter in a pan.
		ours (a)	The person stirred the eggs.
		ours (b)	The person stirred the leek.
		ours (c)	The person stirred the pan with a spatula.
		human	The person placed the cooked leek on a plate.

Table 10: Video description: Example output for a multi-sentence description. We compare our best approach (c), marked in blue with the variants (a),(b), [29], and human descriptions. For details on the approach see Section 6 of the submission and for a discussion of these results see Section C of this pdf. Red underlined marks errors.

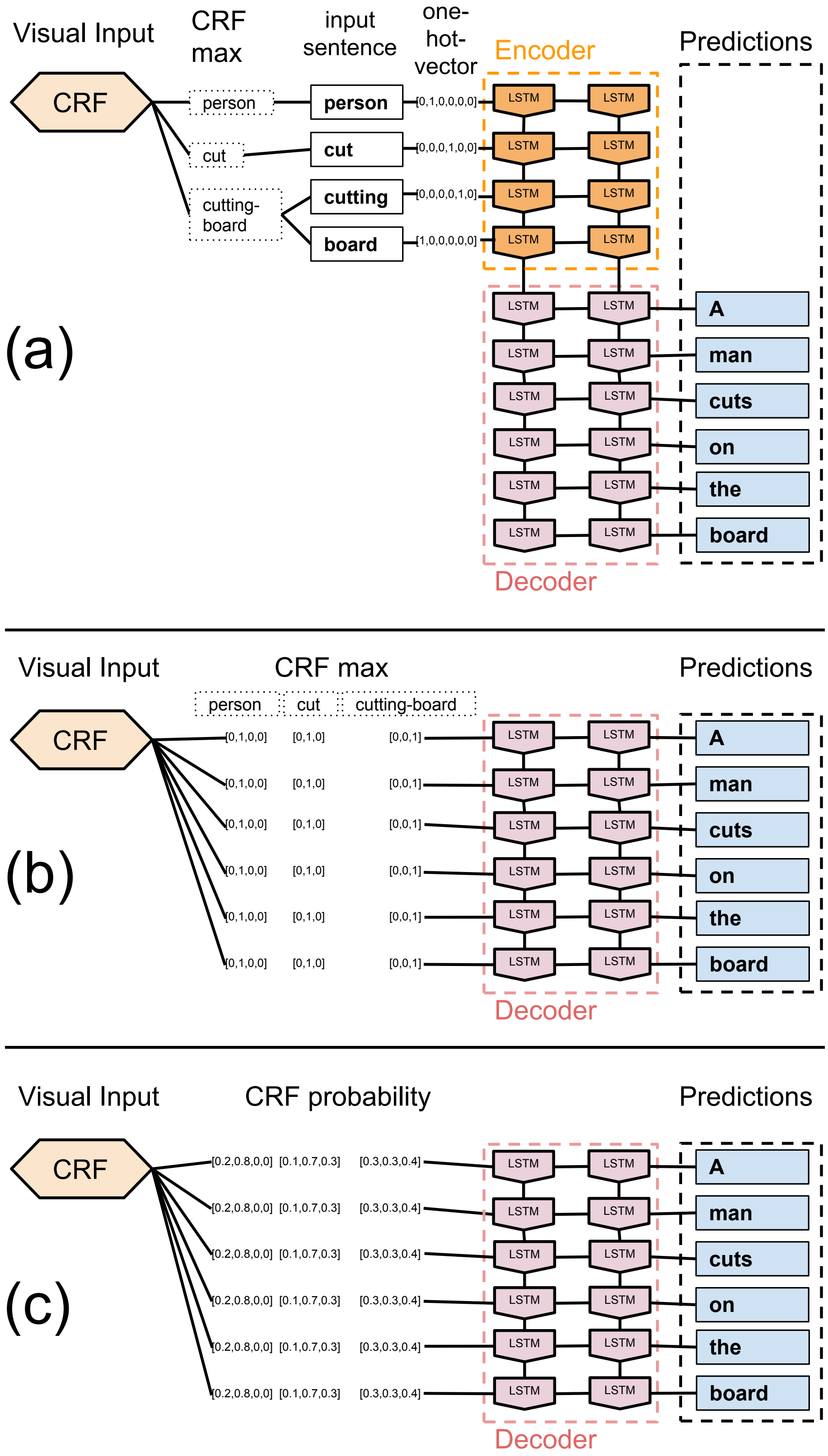


Figure 10: Our approaches to video description. (a) LSTM encoder & decoder with CRF max (b) LSTM decoder with CRF max (c) LSTM decoder with CRF probabilities. (This is a larger and slightly more detailed version of Figure 4 in the submission).