Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson. et al., CVPR, 2018

Presentation Kyunghoon Hur

VISION STUDY

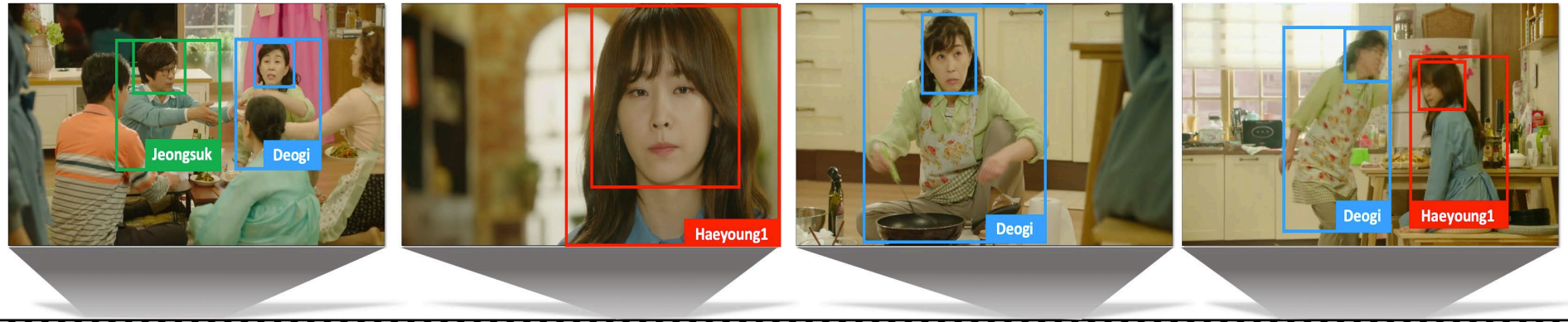
2020.09.15

Background

1 Bottom-Up and Top-Down 2 Attention for Image Captioning 3 and Visual Question Answering 4

Background

Visual Question Answering



Deogi: Mother, have some pancakes
None: Why did you(Deogi) make so much?

Haeyoung1: I(Haeyoung1)'m not getting married.
Deogi: What did you(Haeyoung1) say?

Deogi: You(Haeyoung1) must be out of your mind, saying such things out of the blue.
Haeyoung1: We(Haeyoung1, Taejin) fought planning the wedding.

Difficulty 1

Q : How is Haeyoung1's hair style?
A : Haeyoung1 has a long curly hair.

Difficulty 2

Q : What did Jeongsuk hand over to the man?
A : Jeongsuk handed over a plate to the man.

Difficulty 3

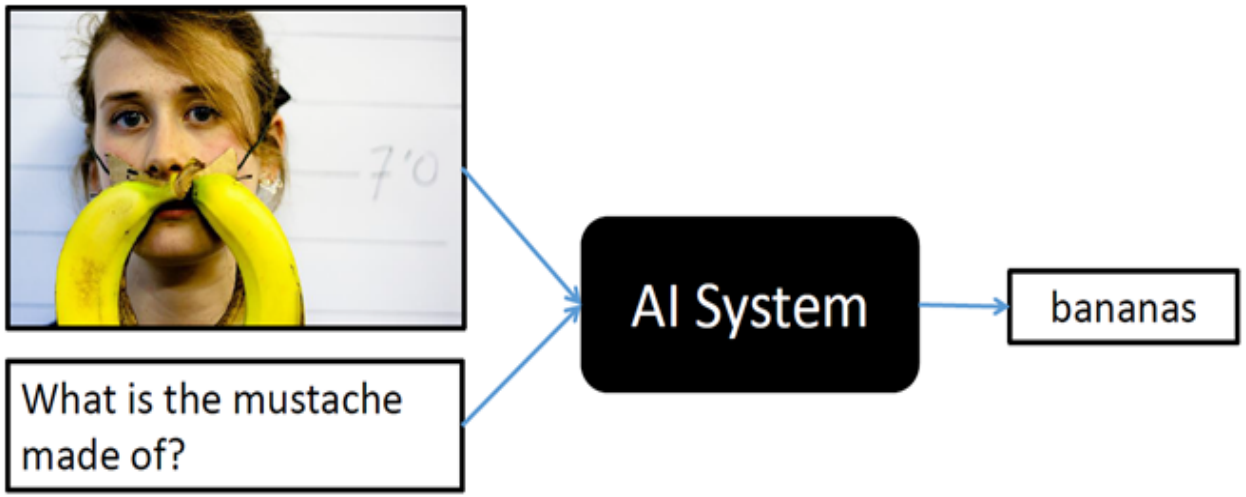
Q : How did Deogi react when Haeyoung1 said Haeyoung1 won't get married?
A : Deogi yelled at Haeyoung1 and hit Haeyoung1's head.

Difficulty 4

Q : Why did Deogi make food a lot?
A : Because Deogi wanted to share the food with her neighborhoods.

Background

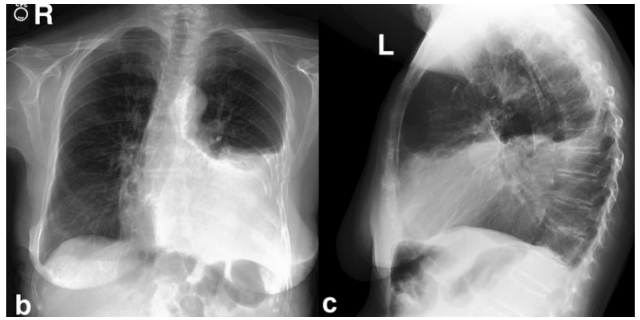
Visual Question Answering



Background

Visual Question Answering

MIMIC-CXR



EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

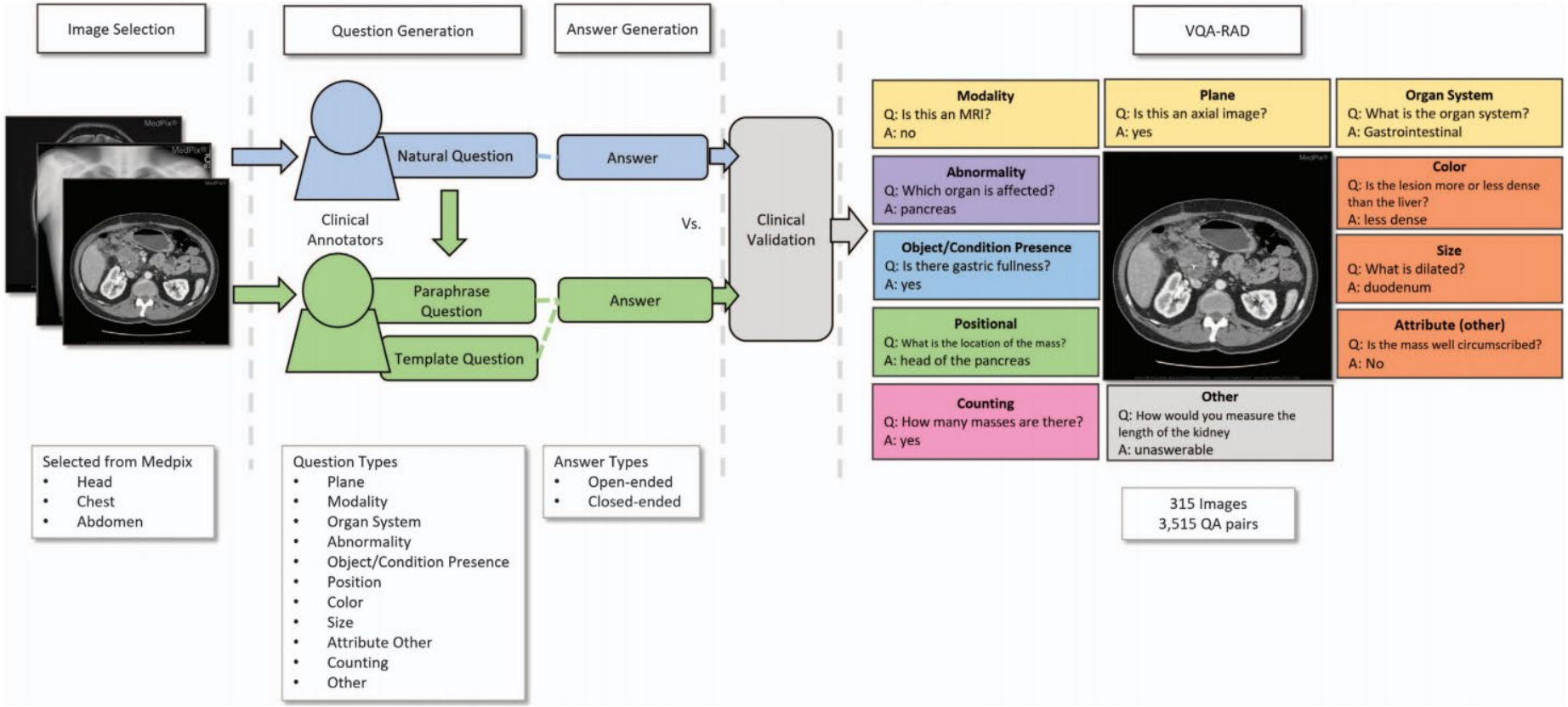
COMPARISON: ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION:

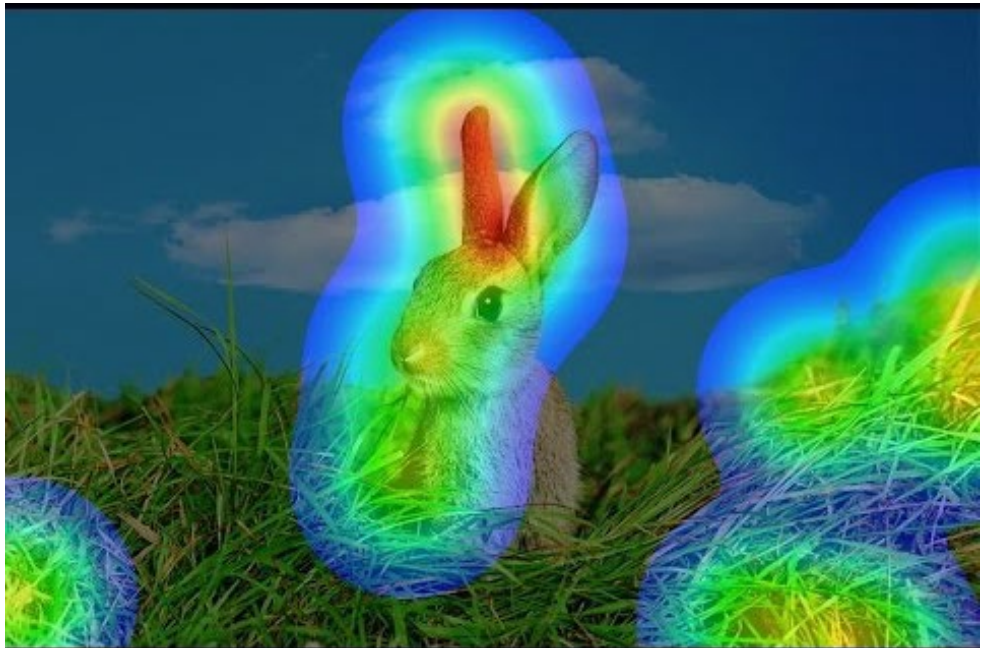
a Large left pleural effusion



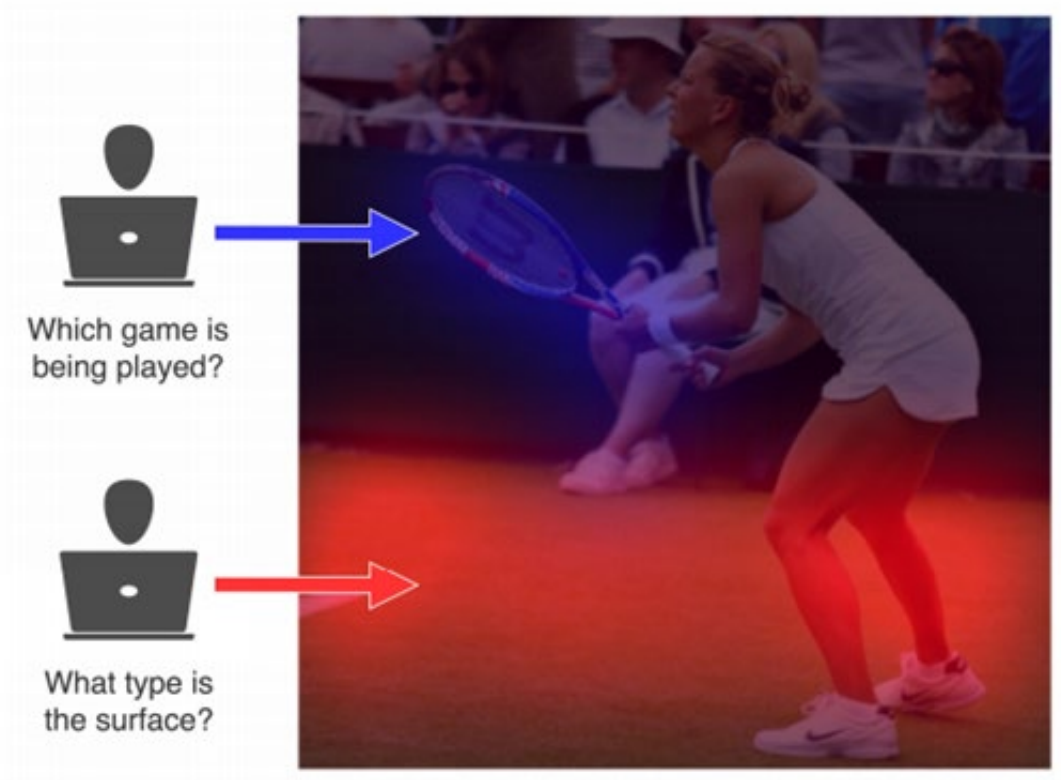
-A dataset of clinically generated visual questions and answers about radiology images (link)

Background

Attention



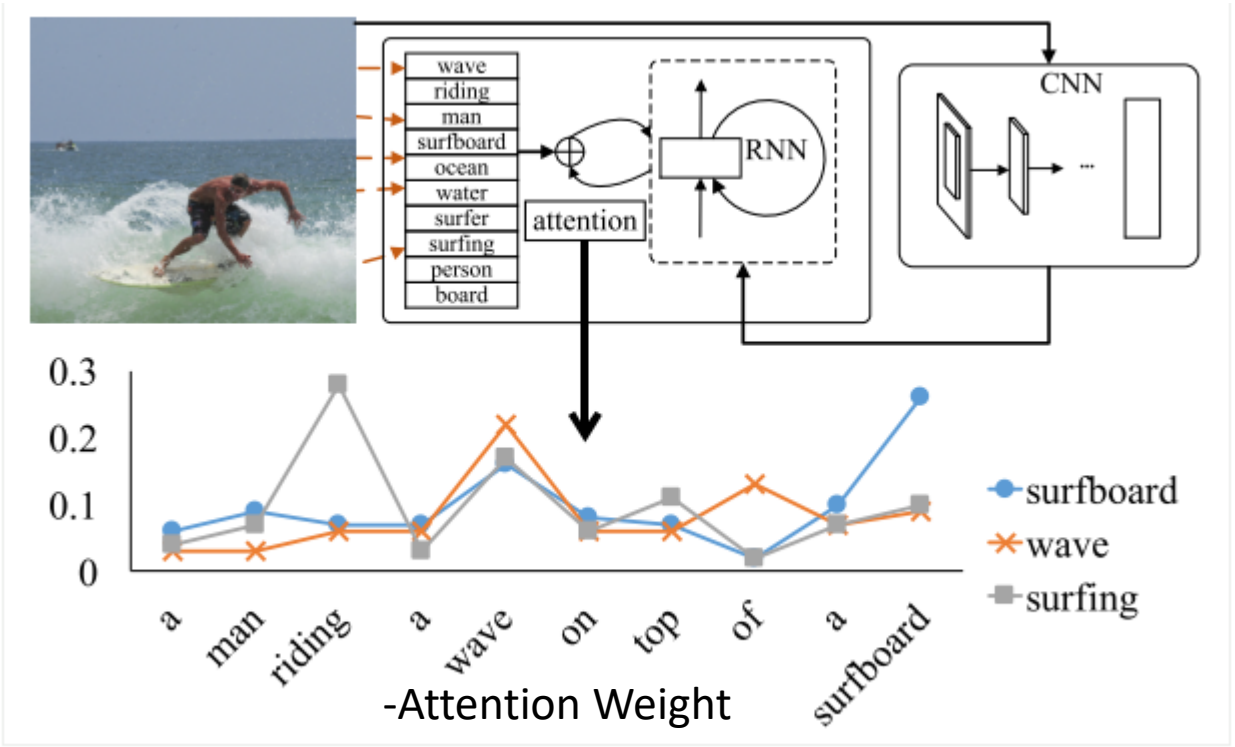
Visual attention



Abhishek Das. et al., EMNLP 2016

Background

Image Captioning

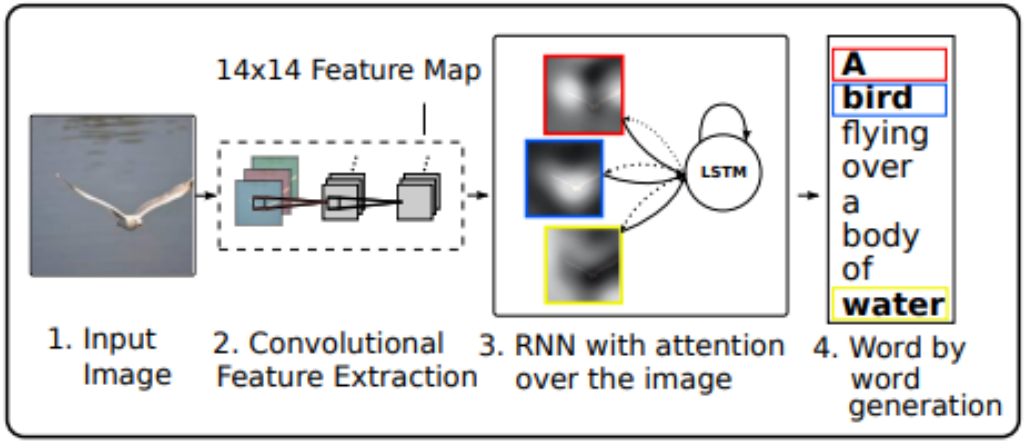


Background

[Bottom-Up and Top-Down]

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu	KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba	JIMMY@PSI.UTORONTO.CA
Ryan Kiros	RKIROS@CS.TORONTO.EDU
Kyunghyun Cho	KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville	AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov	RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel	ZEMEL@CS.TORONTO.EDU
Yoshua Bengio	FIND-ME@THE.WEB



- PMLR(2015)
- Top-down approach attention

Motivation

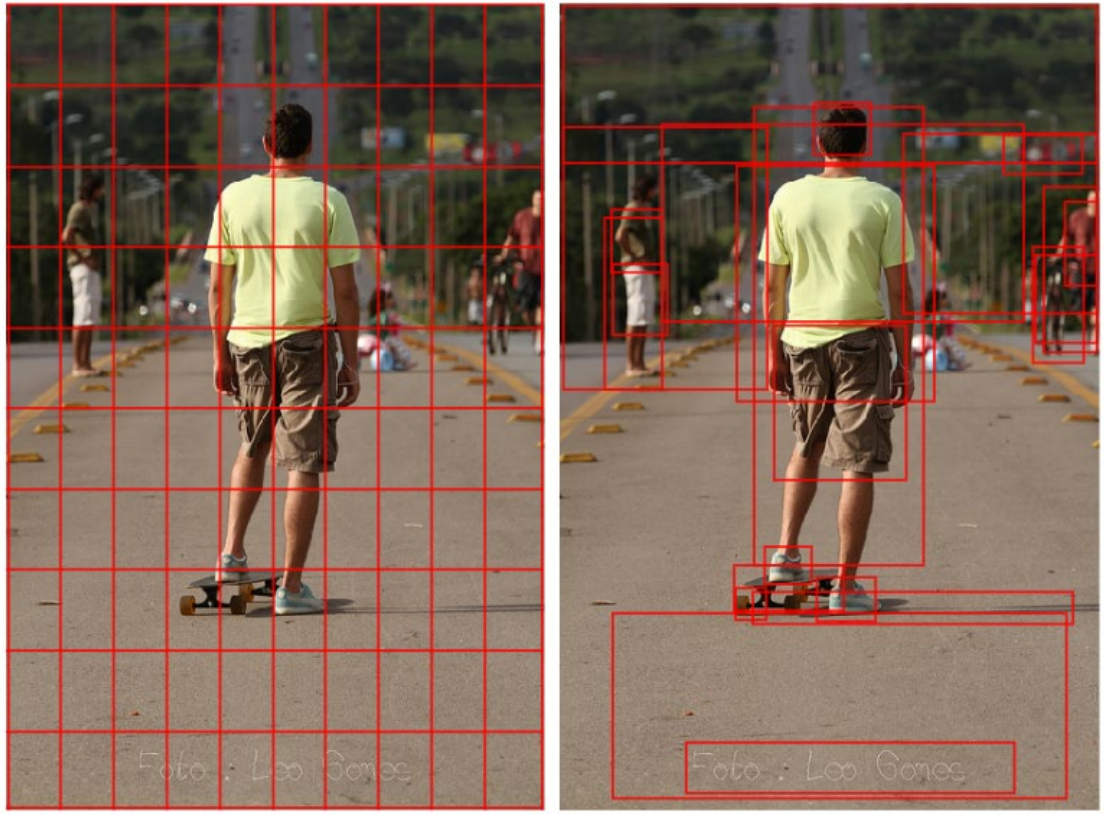
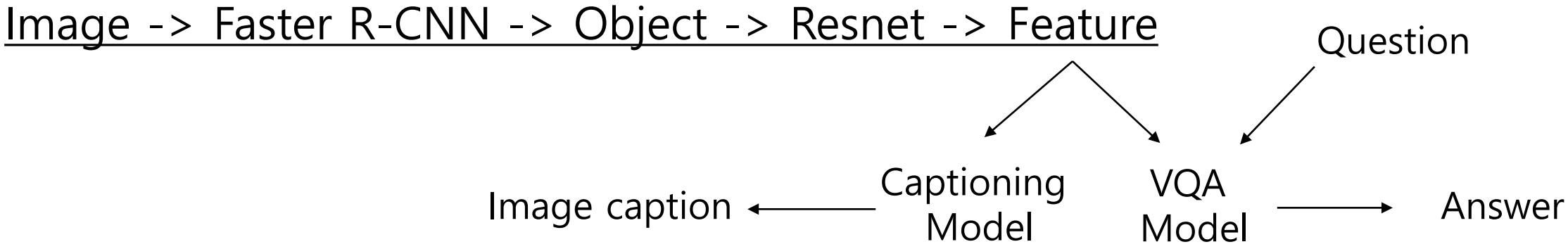


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

Method

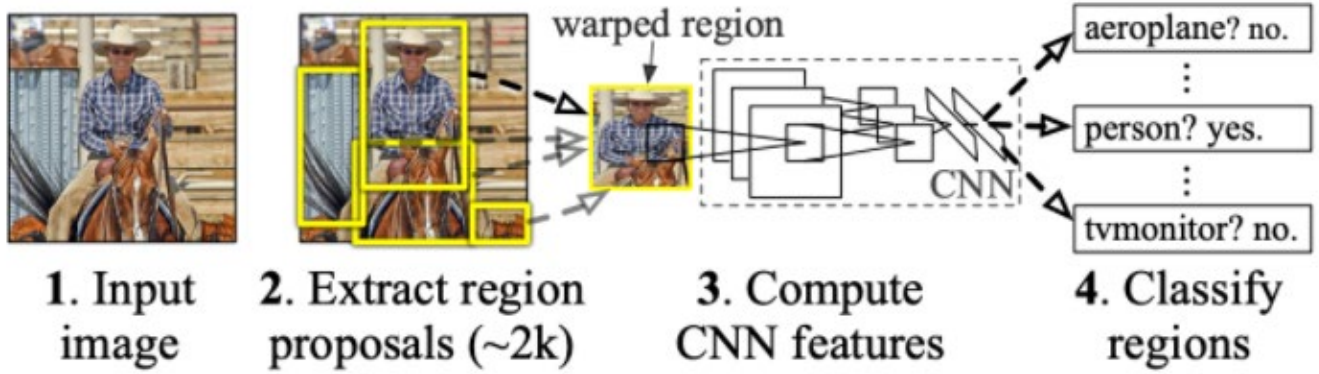
- Bottom-Up Attention Model
- Captioning Model
- VQA Model



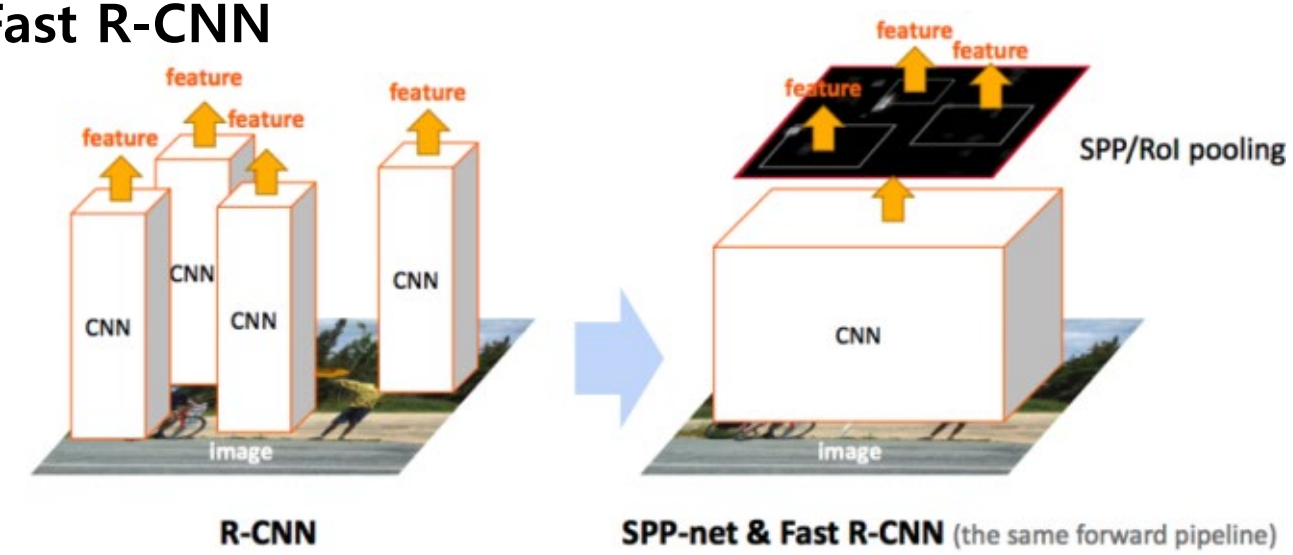
Method

- Bottom-Up Attention Model

1 R-CNN



2 Fast R-CNN



3 Faster R-CNN

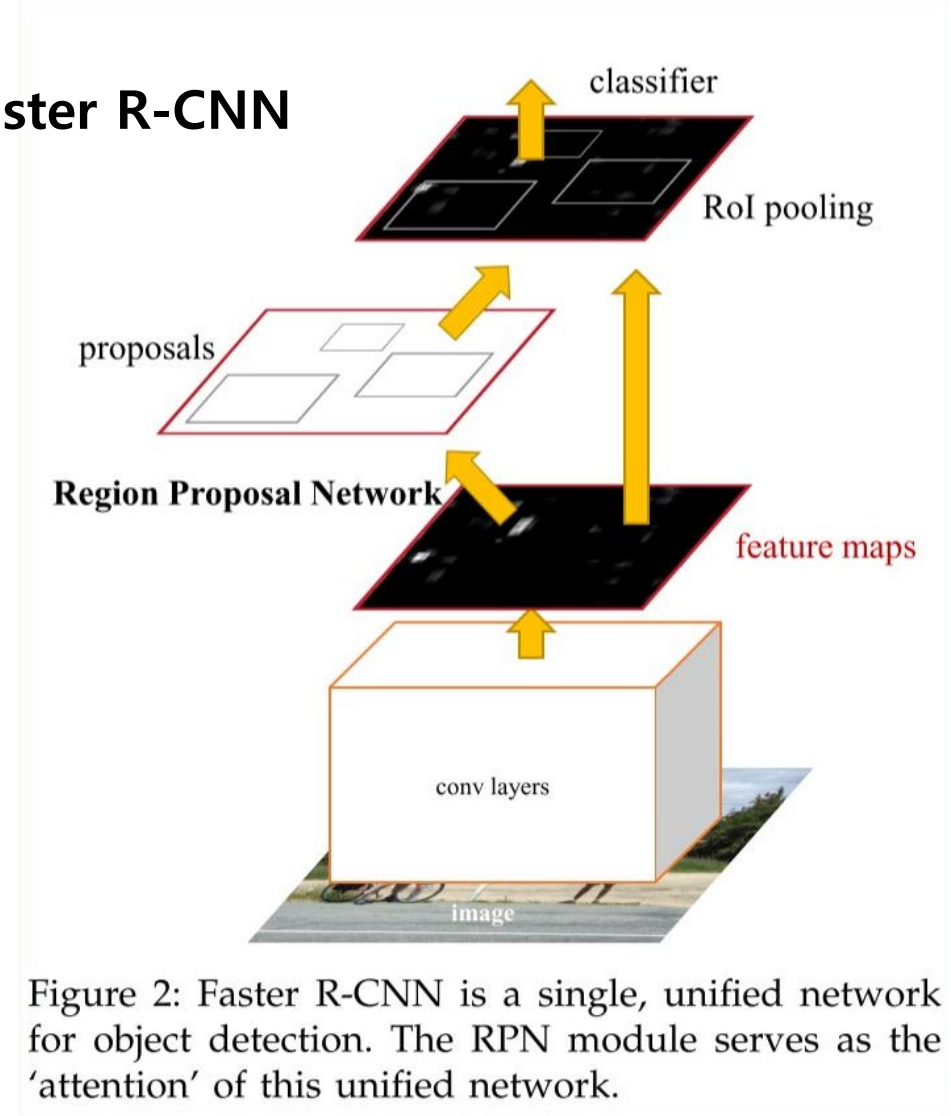
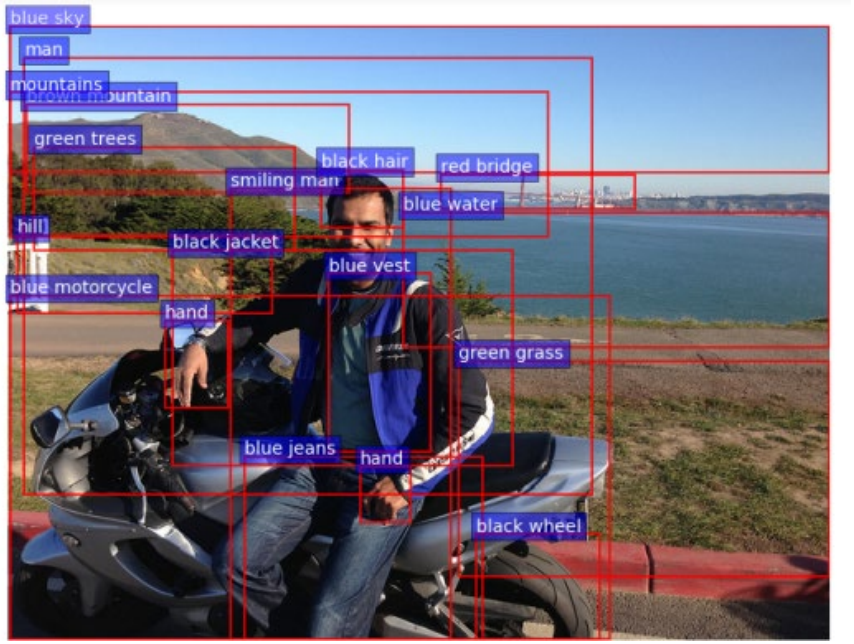
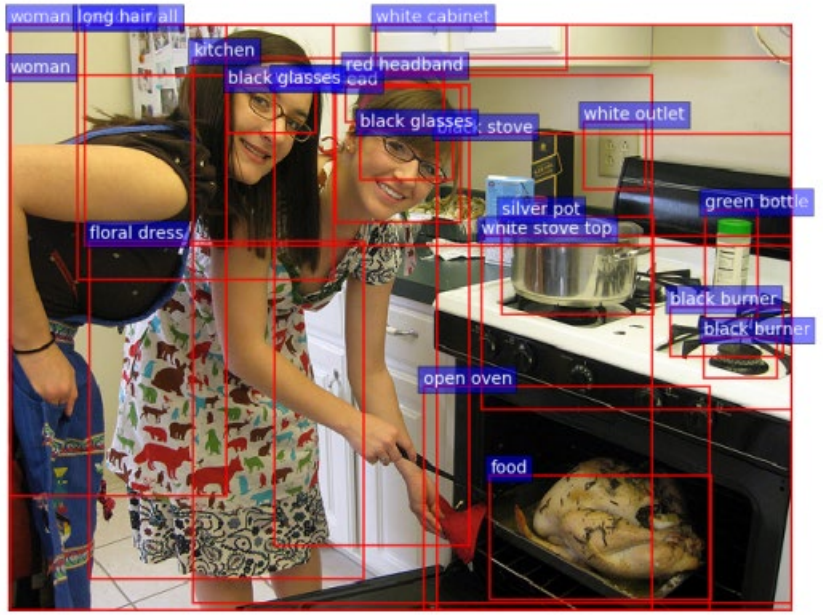


Figure 2: Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.

Method

- Bottom-Up Attention Model



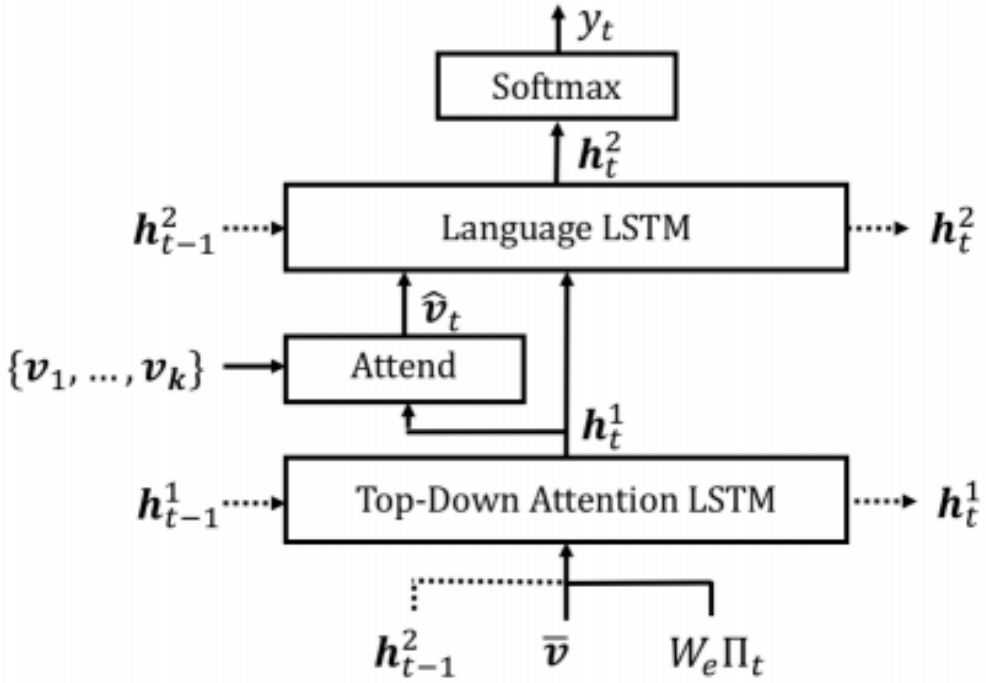


- ✓ Bounding box regression
- ✓ Attribute predictor
- ✓ Region proposal class

- ➢ Output from Faster R-CNN bottom up attention model
- ➢ Each bounding box is labeled with an attribute class
- ➢ VQA and captioning -> utilize only the feature vectors – not the predicted labels

Method

- Captioning Model



$$p(y_t \mid y_{1:t-1}) = \text{softmax}\left(W_p \boldsymbol{h}_t^2 + \boldsymbol{b}_p\right)$$

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t \mid y_{1:t-1})$$

Final output (caption sentence)

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* \mid y_{1:t-1}^*))$$

Caption model loss

$$L_R(\theta) = -\mathbf{E}_{y_{1:T} \sim p_\theta}\left[r(y_{1:T})\right]$$

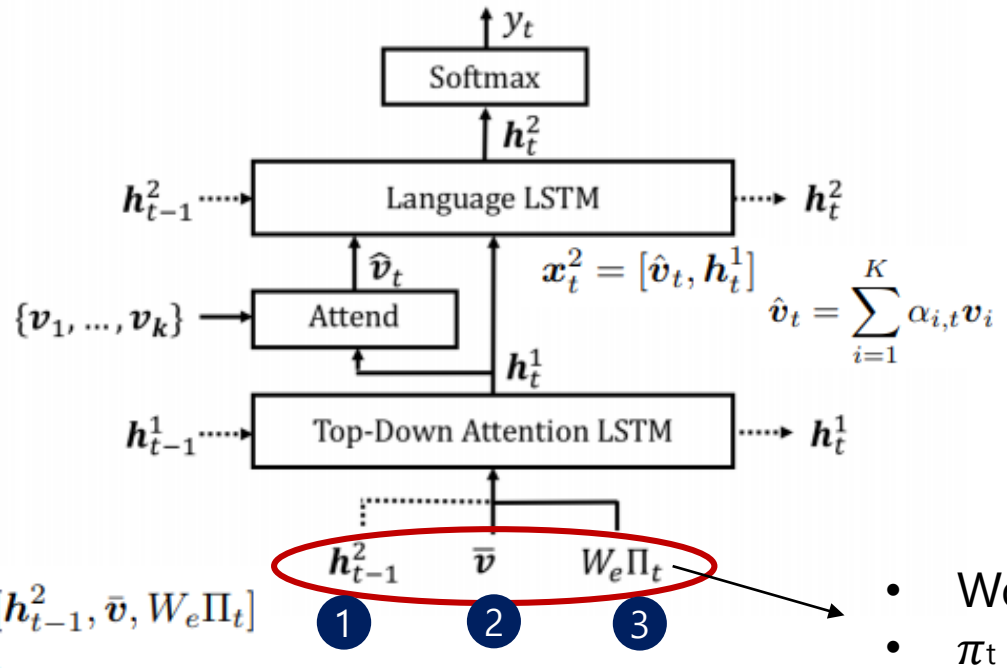
Negative expected score(CIDEr)

$$\nabla_\theta L_R(\theta) \approx -(r(y_{1:T}^s) - r(\hat{y}_{1:T}))\nabla_\theta \log p_\theta(y_{1:T}^s)$$

SCST loss

Method

- Captioning Model



$x_t^1 = [h_{t-1}^2, \bar{v}, W_e \Pi_t]$

$\bar{v} = \frac{1}{k} \sum_i v_i$

- ✓ Two LSTM layers are used to selectively attend to spatial image features
- ✓ Features { v1, …, v_k } -> spatial output of CNN
- ✓ Bottom up + soft top-down attention

- Word embedding matrix for a vocab
- π_t : one-hot encoding of the input word at time step t

1 2 3

- ✓ Input (previous ouput of language LSTM / Image mean pooling feature / previously generated word)

-> h_t (soft-attention) -> bottom up image mean pooling

-> 2nd layer LSTM input - v_t (Language LSTM) generating

Method

- Captioning Model

$$p(y_t \mid y_{1:t-1}) = \text{softmax}\left(W_p \boldsymbol{h}_t^2 + \boldsymbol{b}_p\right)$$

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t \mid y_{1:t-1})$$

Final output (caption sentence)

- ✓ Linear regression soft max
- ✓ Full sentence iterating to final state

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* \mid y_{1:t-1}^*))$$

Caption model loss

$$L_R(\theta) = -\mathbf{E}_{y_{1:T} \sim p_\theta}[r(y_{1:T})]$$

Negative expected score(CIDEr)

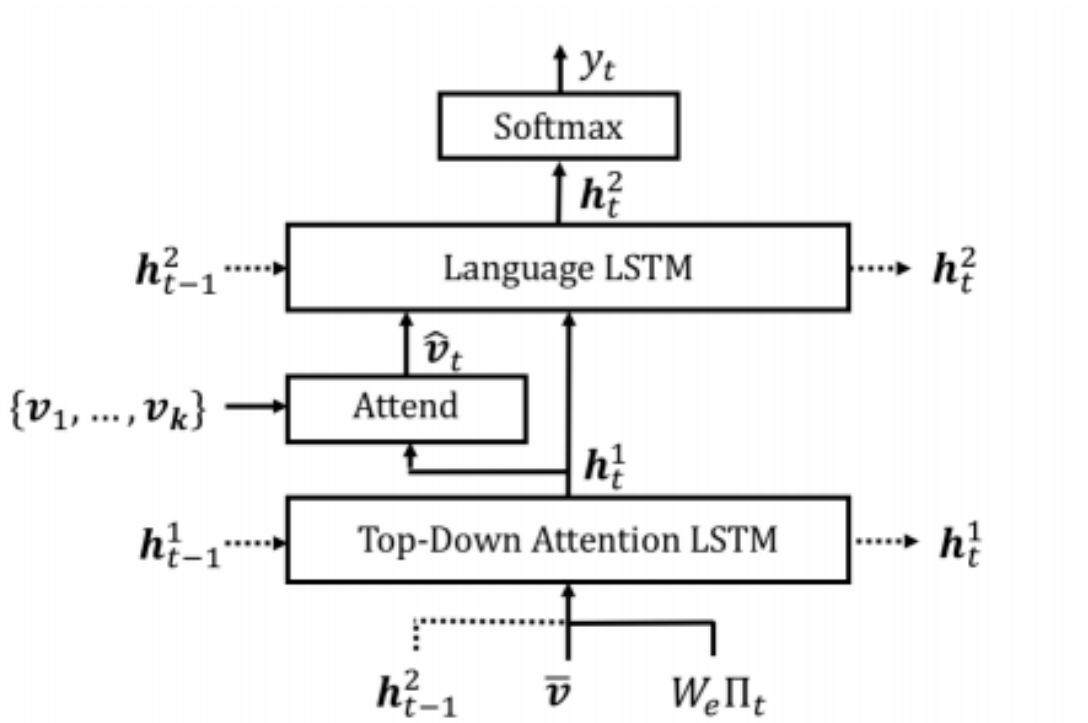
$$\nabla_\theta L_R(\theta) \approx -(r(y_{1:T}^s) - r(\hat{y}_{1:T}))\nabla_\theta \log p_\theta(y_{1:T}^s)$$

SCST loss

- ✓ Captioning model cross entropy loss
- ✓ Fair comparison -> various loss

Method

- Captioning Model



$$p(y_t \mid y_{1:t-1}) = \text{softmax}\left(W_p \boldsymbol{h}_t^2 + \boldsymbol{b}_p\right)$$

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t \mid y_{1:t-1})$$

최종 문장 출력 공식

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^* \mid y_{1:t-1}^*))$$

캡션 모델 loss

$$L_R(\theta) = -\mathbf{E}_{y_{1:T} \sim p_\theta}[r(y_{1:T})]$$

스코어 함수를 반영한 loss

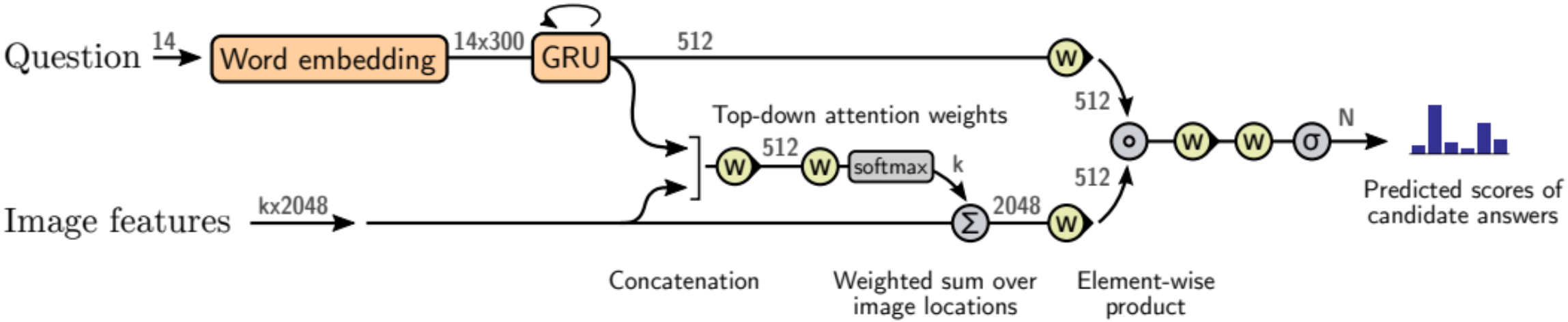
$$\nabla_\theta L_R(\theta) \approx -(r(y_{1:T}^s) - r(\hat{y}_{1:T}))\nabla_\theta \log p_\theta(y_{1:T}^s)$$

SCST 방식 loss

다양한 loss 함수

Method

- VQA Model



Experiment Result

✓ Dataset : MSCOCO 2014 captions dataset & VQA v2.0

	Cross-Entropy Loss						CIDEr Optimization					
	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
SCST:Att2in [33]	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
SCST:Att2all [33]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
Ours: ResNet	74.5	33.4	26.1	54.4	105.4	19.2	76.6	34.0	26.5	54.9	111.1	20.2
Ours: Up-Down	**77.2**	**36.2**	**27.0**	**56.4**	**113.5**	**20.3**	**79.8**	**36.3**	**27.7**	**56.9**	**120.1**	**21.4**
Relative Improvement	4%	8%	3%	4%	8%	6%	4%	7%	5%	4%	8%	6%

Table 1. Single-model image captioning performance on the MSCOCO Karpathy test split. Our baseline ResNet model obtains similar results to SCST [33], the existing state-of-the-art on this test set. Illustrating the contribution of bottom-up attention, our Up-Down model achieves significant (3–8%) relative gains across all metrics regardless of whether cross-entropy loss or CIDEr optimization is used.

Experiment Result

✓ Dataset : MSCOCO & VQA v2.0

- Previous state of the art SCST(Self-critical sequence training)

	Cross-Entropy Loss						CIDEr Optimization					
	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
SCST:Att2in [33]	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
SCST:Att2all [33]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
Ours: ResNet	74.5	33.4	26.1	54.4	105.4	19.2	76.6	34.0	26.5	54.9	111.1	20.2
Ours: Up-Down	**77.2**	**36.2**	**27.0**	**56.4**	**113.5**	**20.3**	**79.8**	**36.3**	**27.7**	**56.9**	**120.1**	**21.4**
Relative Improvement	4%	8%	3%	4%	8%	6%	4%	7%	5%	4%	8%	6%

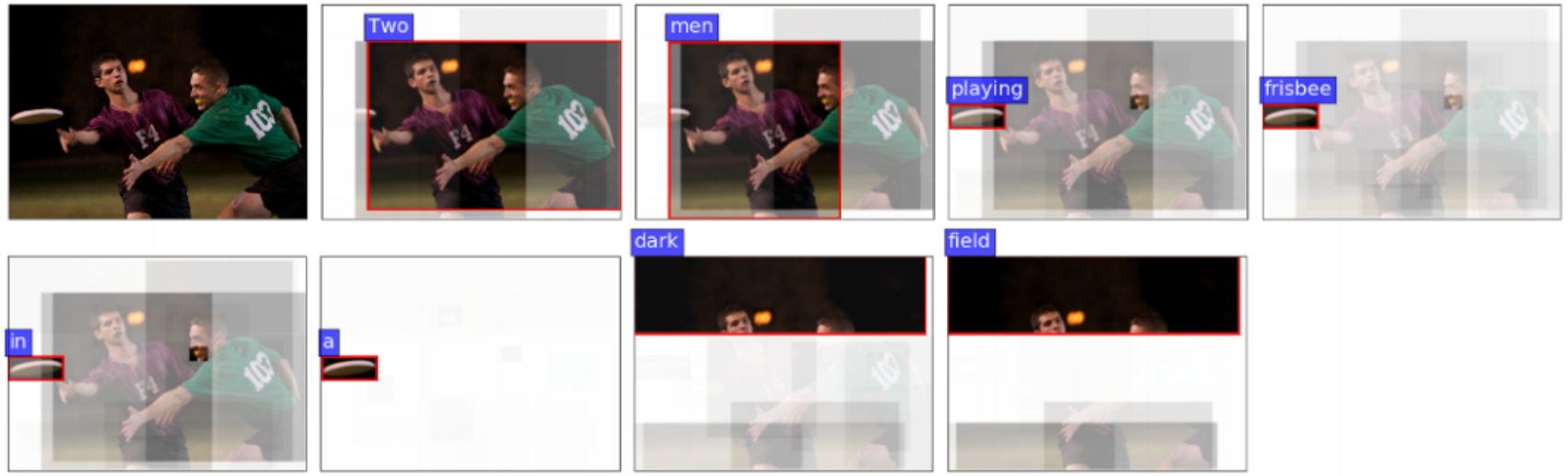
Table 1. Single-model image captioning performance on the MSCOCO Karpathy test split. Our baseline ResNet model obtains similar results to SCST [33], the existing state-of-the-art on this test set. Illustrating the contribution of bottom-up attention, our Up-Down model achieves significant (3–8%) relative gains across all metrics regardless of whether cross-entropy loss or CIDEr optimization is used.

	Cross-Entropy Loss							CIDEr Optimization						
	SPICE	Objects	Attributes	Relations	Color	Count	Size	SPICE	Objects	Attributes	Relations	Color	Count	Size
Ours: ResNet	19.2	35.4	8.6	5.3	12.2	4.1	3.9	20.2	37.0	9.2	6.1	10.6	12.0	**4.3**
Ours: Up-Down	**20.3**	**37.1**	**9.2**	**5.8**	**12.7**	**6.5**	**4.5**	**21.4**	**39.1**	**10.0**	**6.5**	**11.4**	**18.4**	3.2

Experiment Result

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr		SPICE	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Review Net [47]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9	18.5	64.9
Adaptive [27]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9	19.7	67.3
PG-BCMR [24]	75.4	-	59.1	-	44.5	-	33.2	-	25.7	-	55	-	101.3	-	-	-
SCST:Att2all [33]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7	20.7	68.9
LSTM-A_3 [48]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27	35.4	56.4	70.5	116	118	-	-
Ours: Up-Down	**80.2**	**95.2**	**64.1**	**88.8**	**49.1**	**79.4**	**36.9**	**68.5**	**27.6**	**36.7**	**57.1**	**72.4**	**117.9**	**120.5**	**21.5**	**71.5**

Table 3. Highest ranking published image captioning results on the online MSCOCO test server. Our submission, an ensemble of 4 models optimized for CIDEr with different initializations, outperforms previously published work on all reported metrics. At the time of submission (18 July 2017), we also outperformed all unpublished test server submissions.



Two men playing frisbee in a dark field.

Experiment Result

Question: What room are they in? Answer: kitchen

Figure 6. VQA example illustrating attention output. Given the question 'What room are they in?', the model focuses on the stove-top, generating the answer 'kitchen'.

	Yes/No	Number	Other	Overall
Ours: ResNet (1×1)	76.0	36.5	46.8	56.3
Ours: ResNet (14×14)	76.6	36.2	49.5	57.9
Ours: ResNet (7×7)	77.6	37.7	51.5	59.4
Ours: Up-Down	**80.3**	**42.8**	**55.8**	**63.2**
Relative Improvement	3%	14%	8%	6%

Table 4. Single-model performance on the VQA v2.0 validation set. The use of bottom-up attention in the Up-Down model provides a significant improvement over the best ResNet baseline across all question types, even though the ResNet baselines use almost twice as many convolutional layers.

	Yes/No	Number	Other	Overall
d-LSTM+n-I [26, 12]	73.46	35.18	41.83	54.22
MCB [11, 12]	78.82	38.28	53.36	62.27
UPMC-LIP6	82.07	41.06	57.12	65.71
Athena	82.50	44.19	59.97	67.59
HDU-USYD-UNCC	84.50	45.39	59.01	68.09
Ours: Up-Down	**86.60**	**48.64**	**61.15**	**70.34**

Table 5. VQA v2.0 test-standard server accuracy as at 8 August 2017, ranking our submission against published and unpublished work for each question type. Our approach, an ensemble of 30 models, outperforms all other leaderboard entries.

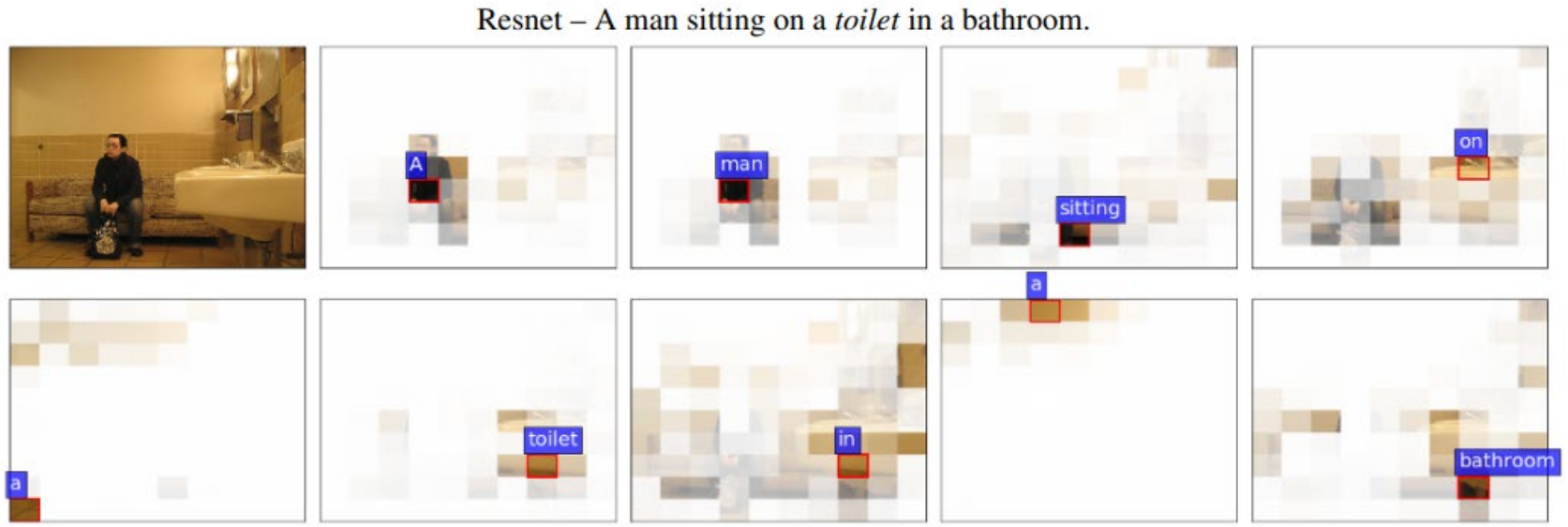
Conclusion & Discussion

- Previous Top-down -> Bottom-up + Top-down
- Attention more naturally involved in task (Image captioning & VQA)
- Follow-up papers cites bottom-up attention approach
- Attention / Object detection / Captioning model / VQA model

End of presentation

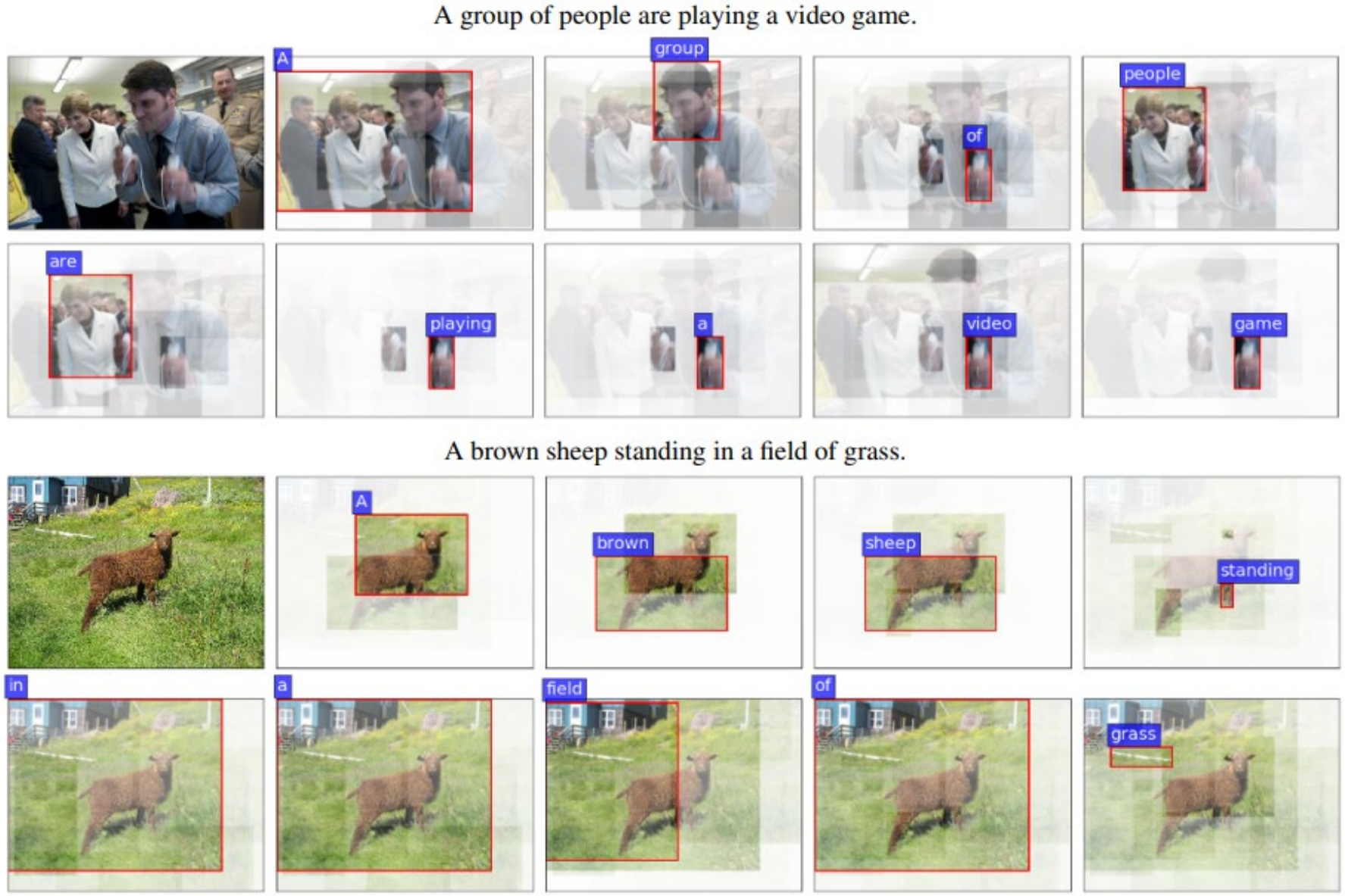
Thank you!

Appendix


Resnet – A man sitting on a *toilet* in a bathroom.
A
man
sitting
on
a
toilet
in
bathroom
a


Up-Down – A man sitting on a *couch* in a bathroom.
A
man
sitting
on
a
couch
in
a
bathroom

Appendix


A group of people are playing a video game.
A
group
people
of
are
playing
a
video
game
A brown sheep standing in a field of grass.
A
brown
sheep
standing
in
a
field
of
grass

Appendix

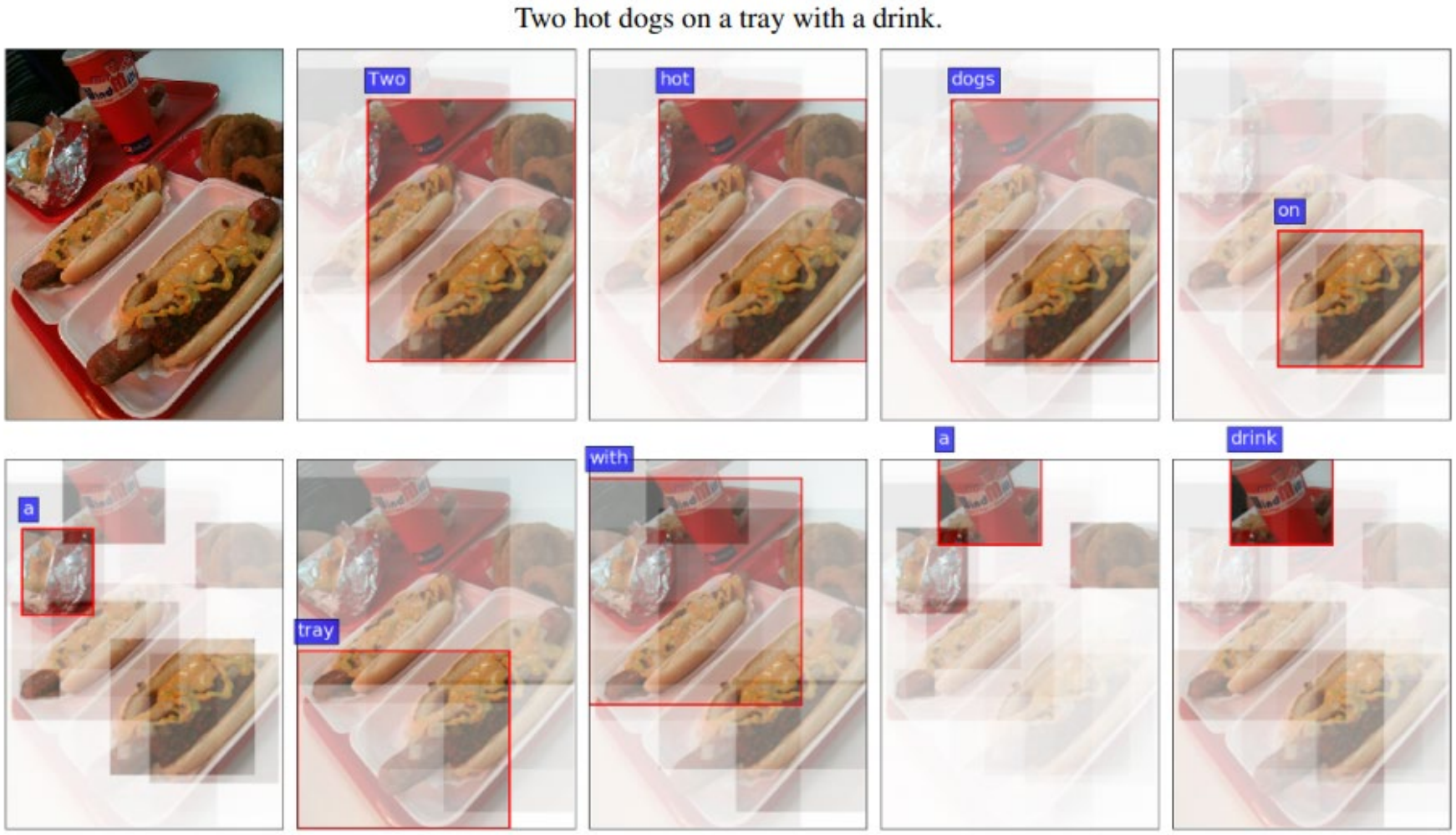
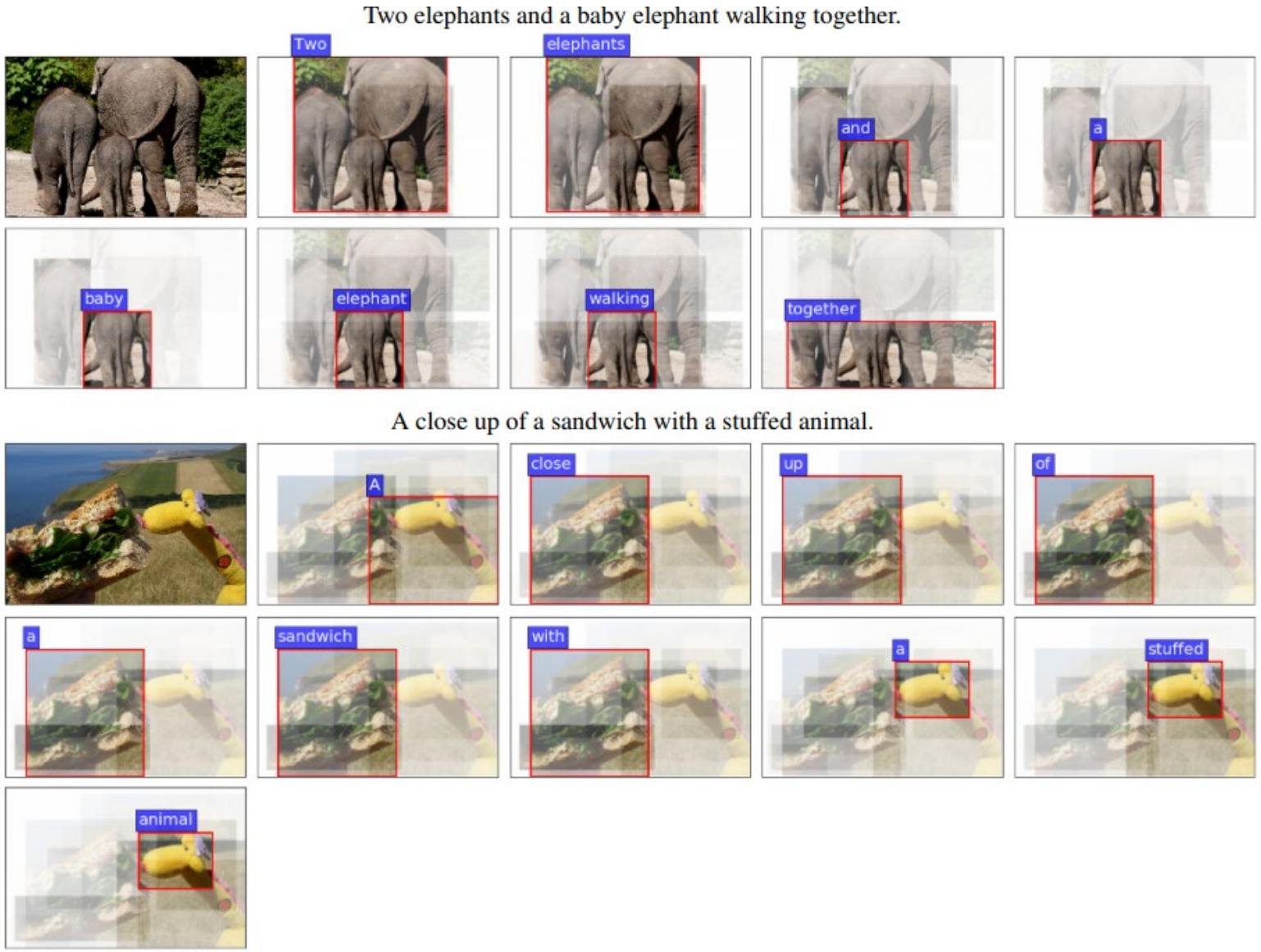


Figure 8. Examples of generated captions showing attended image regions. Attention is given to fine details, such as: (1) the man's hands holding the game controllers in the top image, and (2) the sheep's legs when generating the word 'standing' in the middle image. Our approach can avoid the trade-off between coarse and fine levels of detail.

Appendix


Two elephants and a baby elephant walking together.
Two
elephants
and
a
baby
elephant
walking
together
A close up of a sandwich with a stuffed animal.
A
close
up
of
a
sandwich
with
a
stuffed
animal