simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions

Fenglin Liu[1], Xuancheng Ren[2]*, Yuanxin Liu[1], Houfeng Wang[2] and Xu Sun[2]

[1] Beijing University of Posts and Telecommunications, Beijing, China

[2] Peking University, Beijing, China

* Equal Contributions

北京郵電大學 Beijing University of Posts and Telecommunications x 北京大學 PEKING UNIVERSITY

CONTENTS

北京郵電大學 Beijing University of Posts and Telecommunications X 北京大學 PEKING UNIVERSITY

1 Introduction

simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions

Soft-Attention: a open laptop computer sitting on top of a table

ATT-FCN: a dog sitting on a desk with a laptop computer and mouse

simNet: a open laptop computer and mouse sitting on a table with a dog nearby

Figure 1: Examples of using different attention mechanisms.

·**Soft-Attention:** Show, attend and tell: Neural image caption generation with visual attention. In PMLR 2015

·**ATT-FCN :** Image captioning with semantic attention. In CVPR 2016

Introduction: Soft-Attention

Soft-Attention: a open laptop computer sitting on top of a table

ATT-FCN: a dog sitting on a desk with a laptop computer and mouse

simNet: a open laptop computer and mouse sitting on a table with a dog nearby

→ omitting "dog" and "mouse"

Soft-Attention: Image —encode→ Image Features —decode→ Caption

·**Soft-Attention:** Show, attend and tell: Neural image caption generation with visual attention. In PMLR 2015

Introduction: ATT-FCN

Soft-Attention: a open laptop computer sitting on top of a table

ATT-FCN: a dog sitting on a desk with a laptop computer and mouse → missing "open" and mislocating "dog"

simNet: a open laptop computer and mouse sitting on a table with a dog nearby

ATT-FCN: Image —encode→ Image Keywords —decode→ Caption

·**ATT-FCN :** Image captioning with semantic attention. In CVPR 2016

Introduction: SimNet

Soft-Attention: a open laptop computer sitting on top of a table

ATT-FCN: a dog sitting on a desk with a laptop computer and mouse

simNet: a open laptop computer and mouse sitting on a table with a dog nearby

Introduction: Main idea

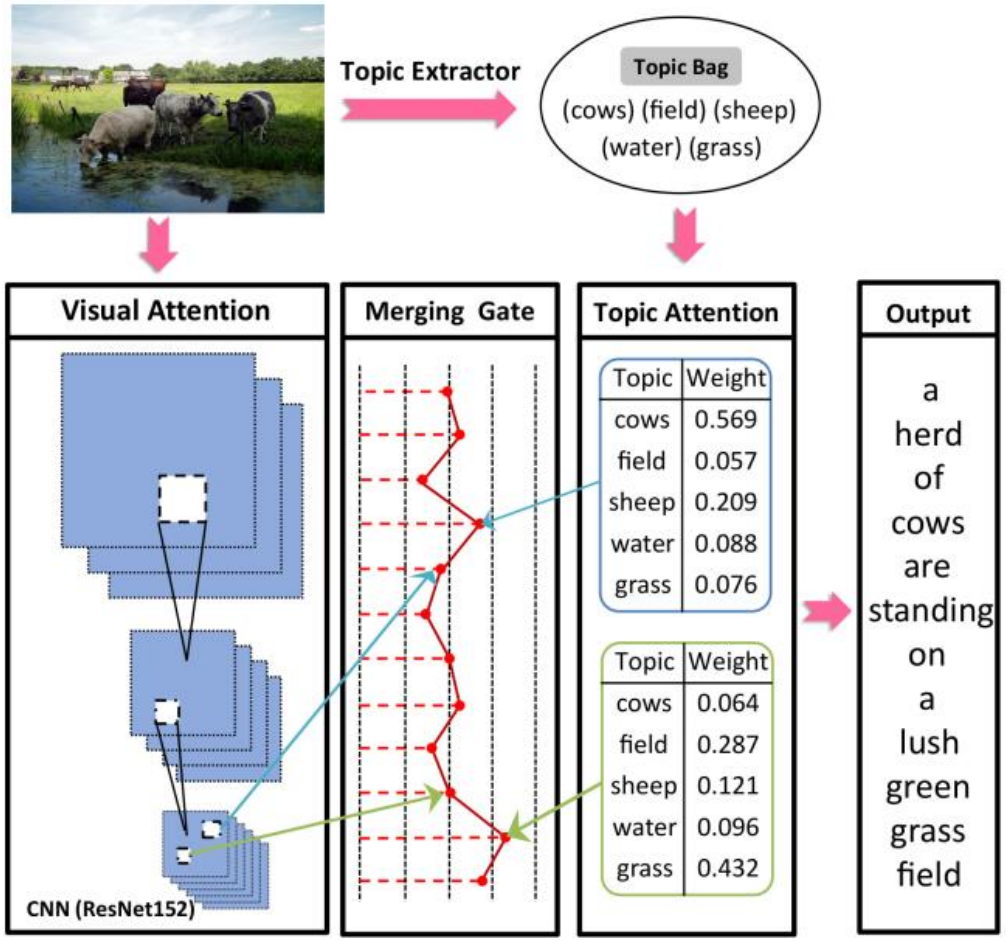


Figure 2: Illustration of the main idea.

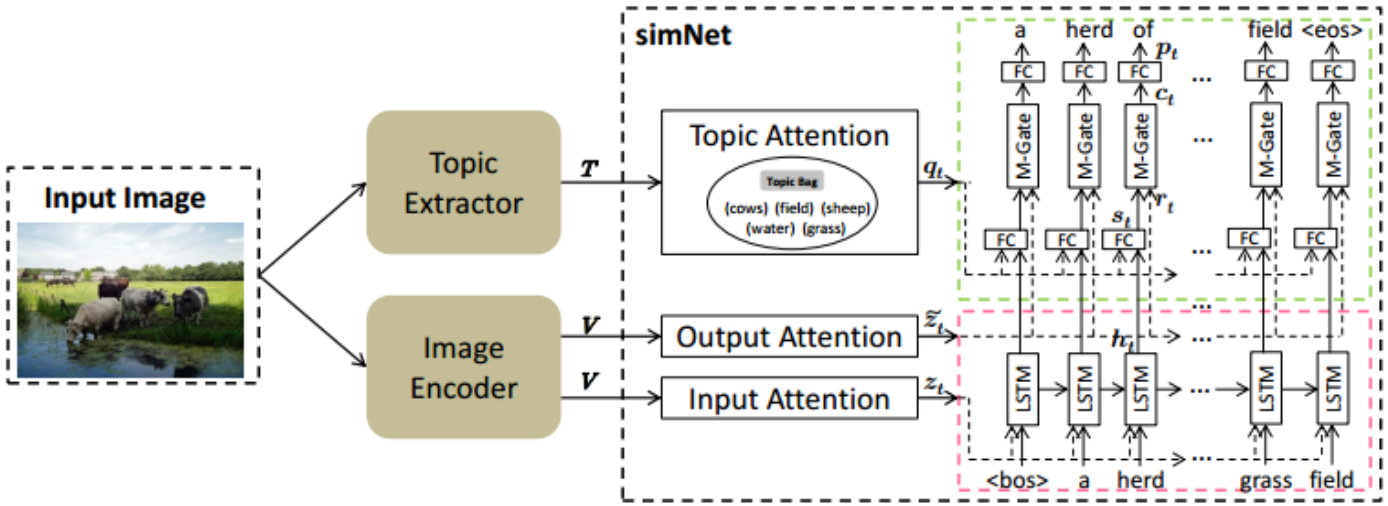
- The visual information captured by CNN
- The topics extracted by a topic extractor
- The merging gate then adaptively adjusts the weight between visual attention and topic attention

Contributions

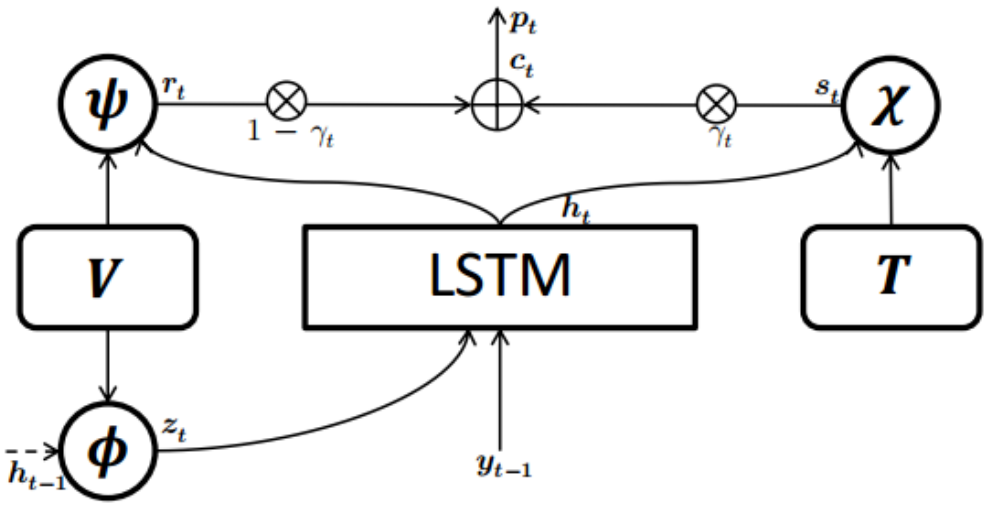
- We propose a novel approach that can effectively merge the information in the image and the topics.
- The generated captions are both detailed and comprehensive.
- The proposed approach outperforms previous works in terms of SPICE, which correlates the best with human judgments.

Approach

Overview



(a) The overall framework.



(b) The data flow in the proposed simNet.

Figure 3: Illustration of the proposed approach. In the right plot, we use ϕ, ψ, χ to denote input attention, output attention, and topic attention, respectively.

Approach: Image Encoder

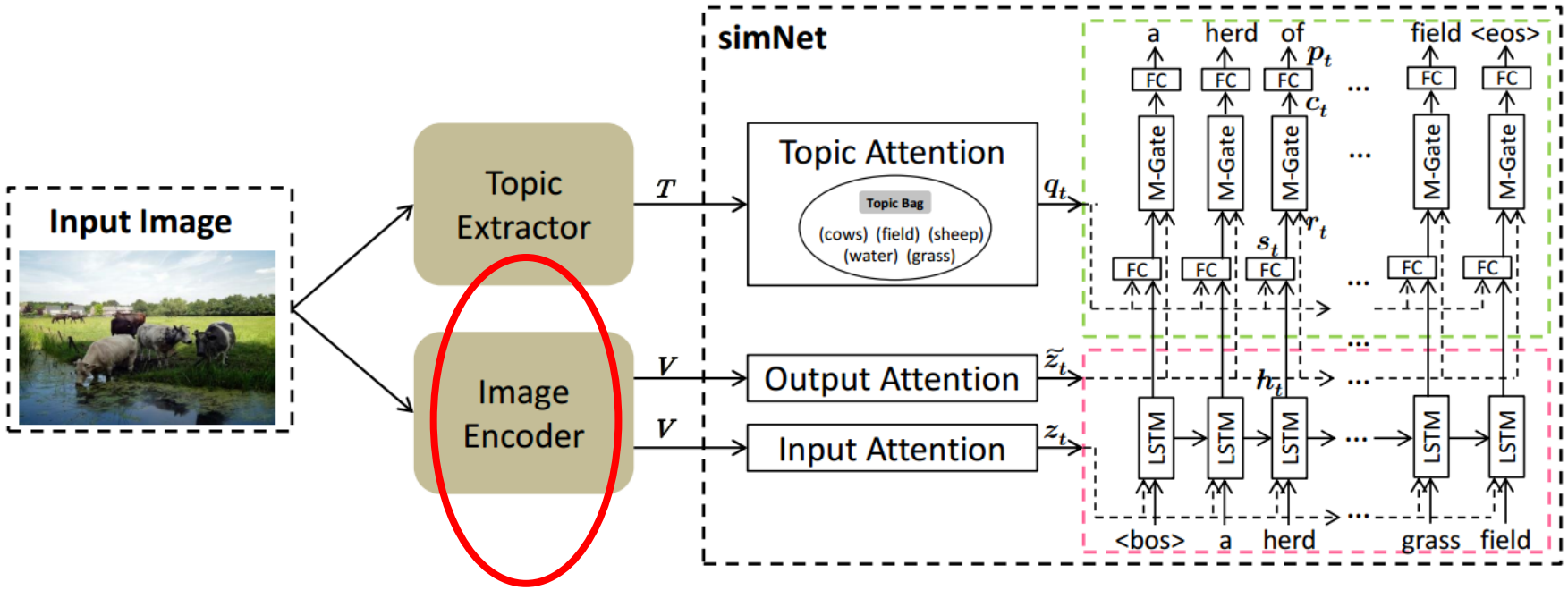
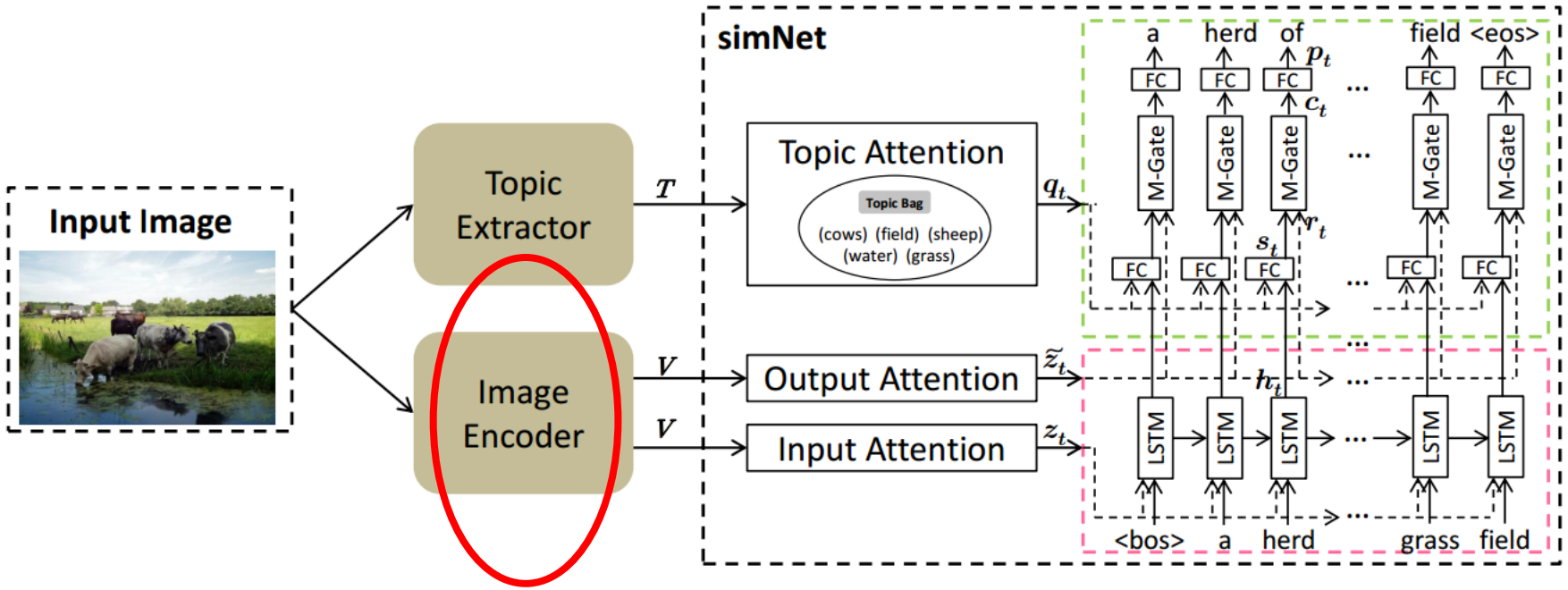


Image Encoder: ResNet152

He et al., 2016: Deep residual learning for image recognition. In CVPR 2016.

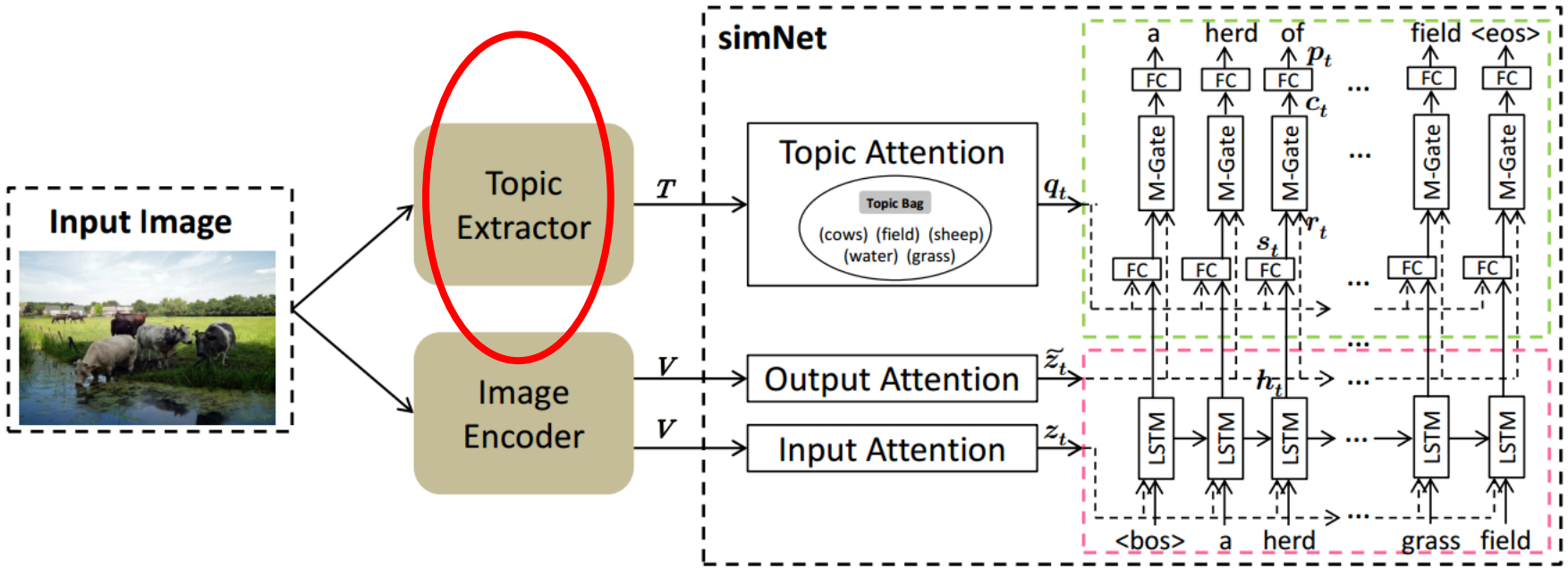
Approach: Image Encoder



Feature map: $\boldsymbol{V} = \boldsymbol{W}^{V,I}\mathrm{CNN}(\boldsymbol{I})$ (1)

where $\boldsymbol{I}$ is the input image, and $\boldsymbol{W}^{V,I}$ shrinks the last dimension of the output.

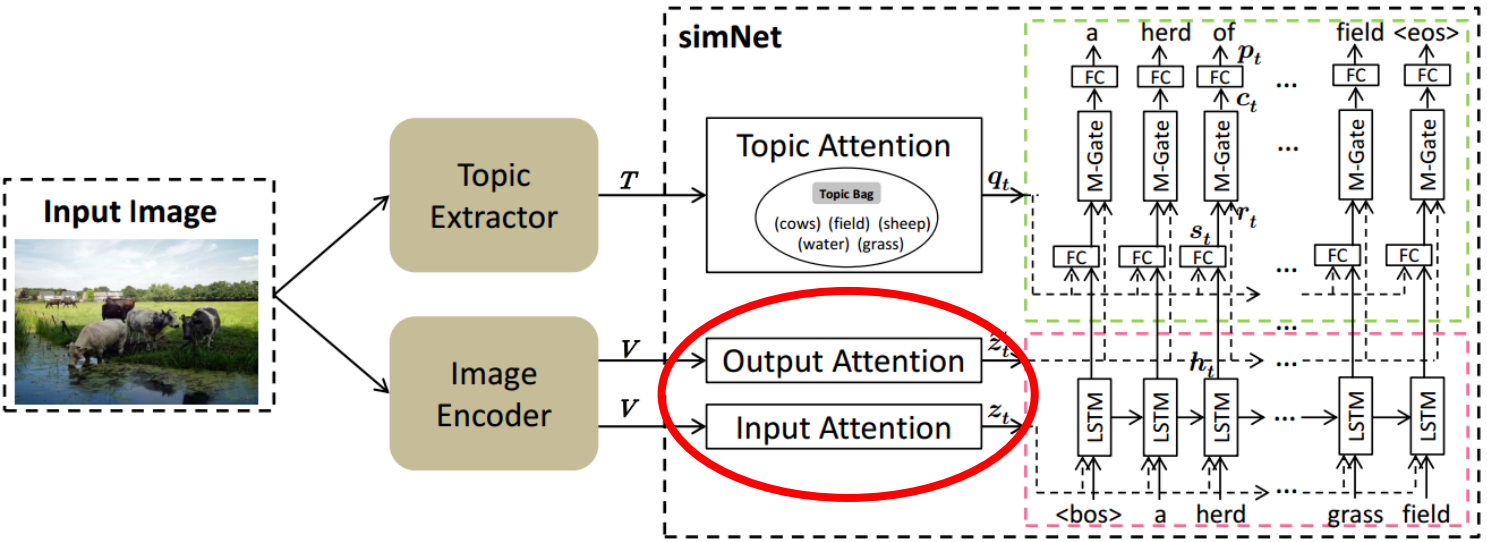
Approach: Topic Extractor



Topic Extractor: Multiple Instance Learning

Zhang et al., 2006: Multiple instance boosting for object detection. In NIPS 2006.
Fang et al., 2015: From captions to visual concepts and back. In CVPR2015

Approach: Input Attention



Input Attention:

$$\boldsymbol{Z}_t = \tanh(\boldsymbol{W}^{Z,V}\boldsymbol{V} \oplus \boldsymbol{W}^{Z,h}\boldsymbol{h}_{t-1}) \quad (2)$$

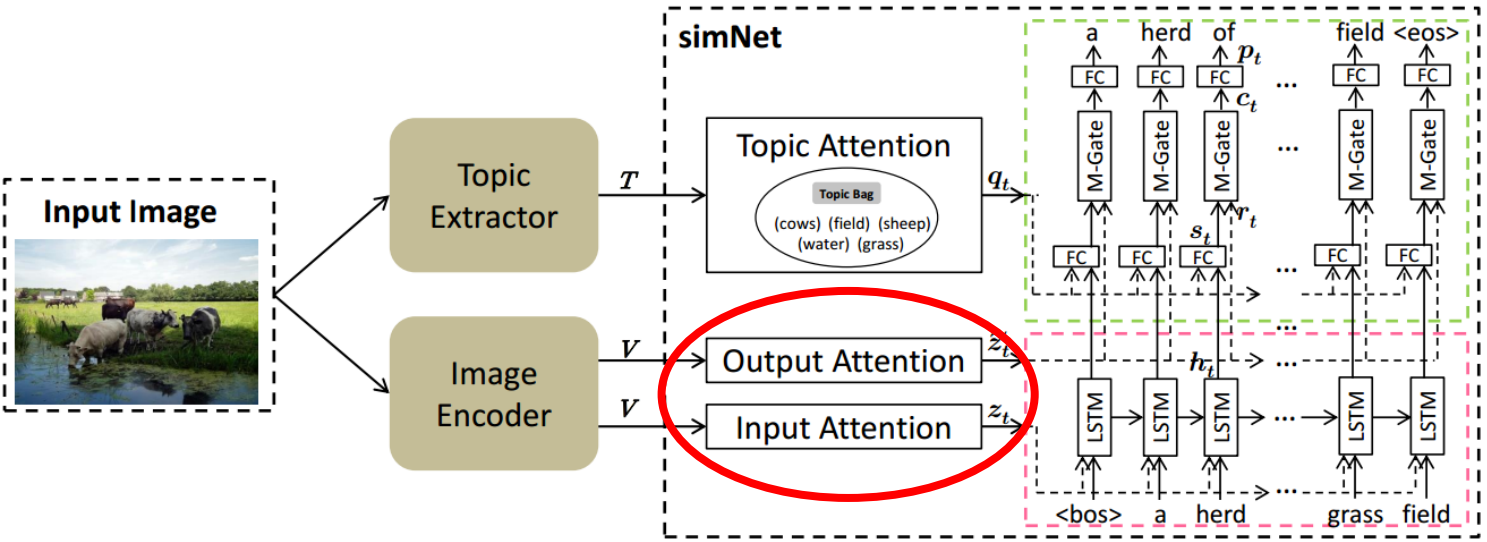
$$\boldsymbol{\alpha}_t = \mathrm{softmax}(\boldsymbol{Z}_t\boldsymbol{w}^{\alpha,Z}) \quad (3)$$

$$\boldsymbol{z}_t = \boldsymbol{V}\boldsymbol{\alpha}_t \quad (4)$$

$$\boldsymbol{h}_t = \mathrm{LSTM}\left(\begin{bmatrix}\boldsymbol{z}_t \\ \boldsymbol{y}_{t-1}\end{bmatrix}, \boldsymbol{h}_{t-1}\right) \quad (5)$$

Xu et al., 2015 : Show, attend and tell: Neural image caption generation with visual attention. In PMLR 2015

Approach: Output Attention



Output Attention:

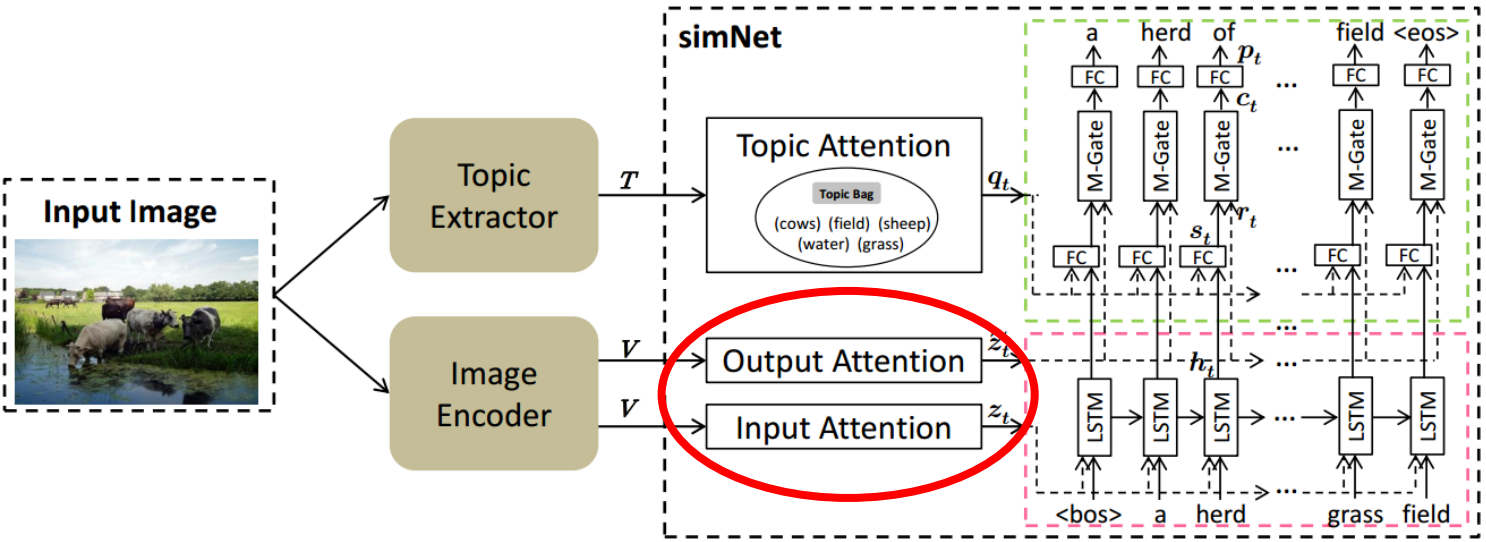
$$\widetilde{\boldsymbol{Z}}_t = \tanh(\widetilde{\boldsymbol{W}}^{Z,V}\boldsymbol{V} \oplus \widetilde{\boldsymbol{W}}^{Z,h}\boldsymbol{h}_t) \quad (6)$$

$$\widetilde{\boldsymbol{\alpha}}_t = \mathrm{softmax}(\widetilde{\boldsymbol{Z}}_t\widetilde{\boldsymbol{w}}^{\alpha,Z}) \quad (7)$$

$$\widetilde{\boldsymbol{z}}_t = \boldsymbol{V}\widetilde{\boldsymbol{\alpha}}_t \quad (8)$$

You et al., 2016 : Image captioning with semantic attention. In CVPR 2016
Lu et al ., 2017 : Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In CVPR 2017

Approach: Visual Information



Output Attention:

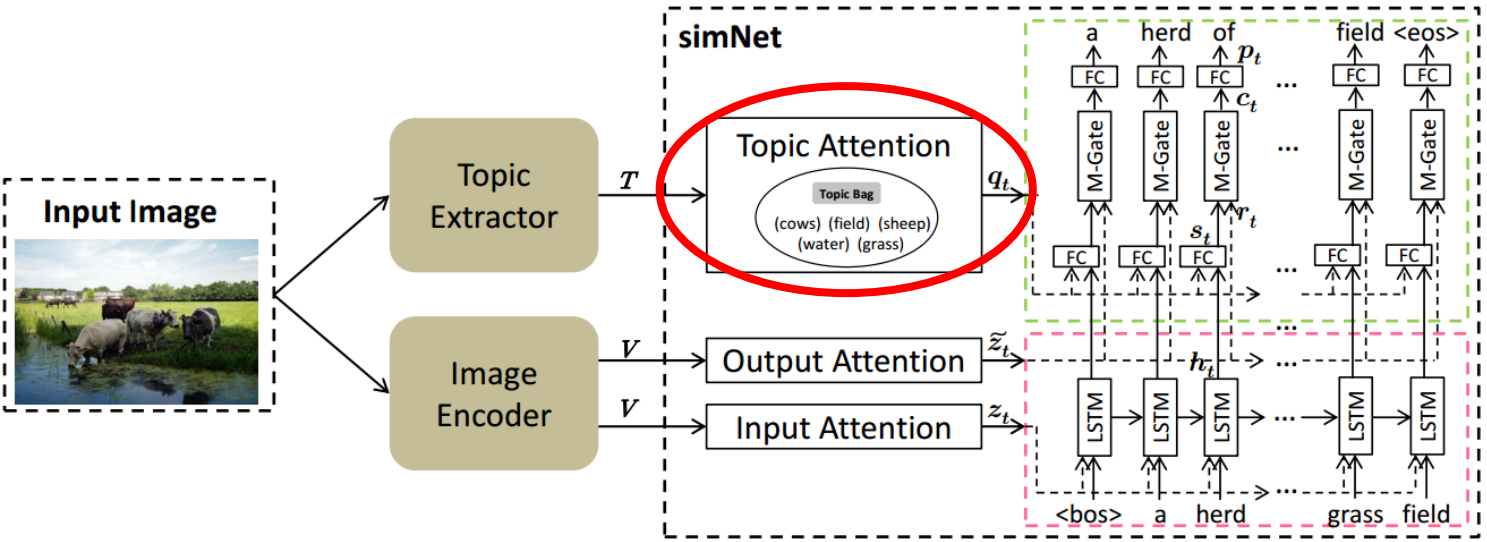
$$\widetilde{\boldsymbol{Z}}_t = \tanh(\widetilde{\boldsymbol{W}}^{Z,V}\boldsymbol{V} \oplus \widetilde{\boldsymbol{W}}^{Z,h}\boldsymbol{h}_t) \quad (6)$$

$$\widetilde{\boldsymbol{\alpha}}_t = \mathrm{softmax}(\widetilde{\boldsymbol{Z}}_t\widetilde{\boldsymbol{w}}^{\alpha,Z}) \quad (7)$$

$$\widetilde{\boldsymbol{z}}_t = \boldsymbol{V}\widetilde{\boldsymbol{\alpha}}_t \quad (8)$$

the visual information: $\boldsymbol{r}_t = \tanh(\boldsymbol{W}^{s,z}\widetilde{\boldsymbol{z}}_t)$

Approach: Previous Topic Attention



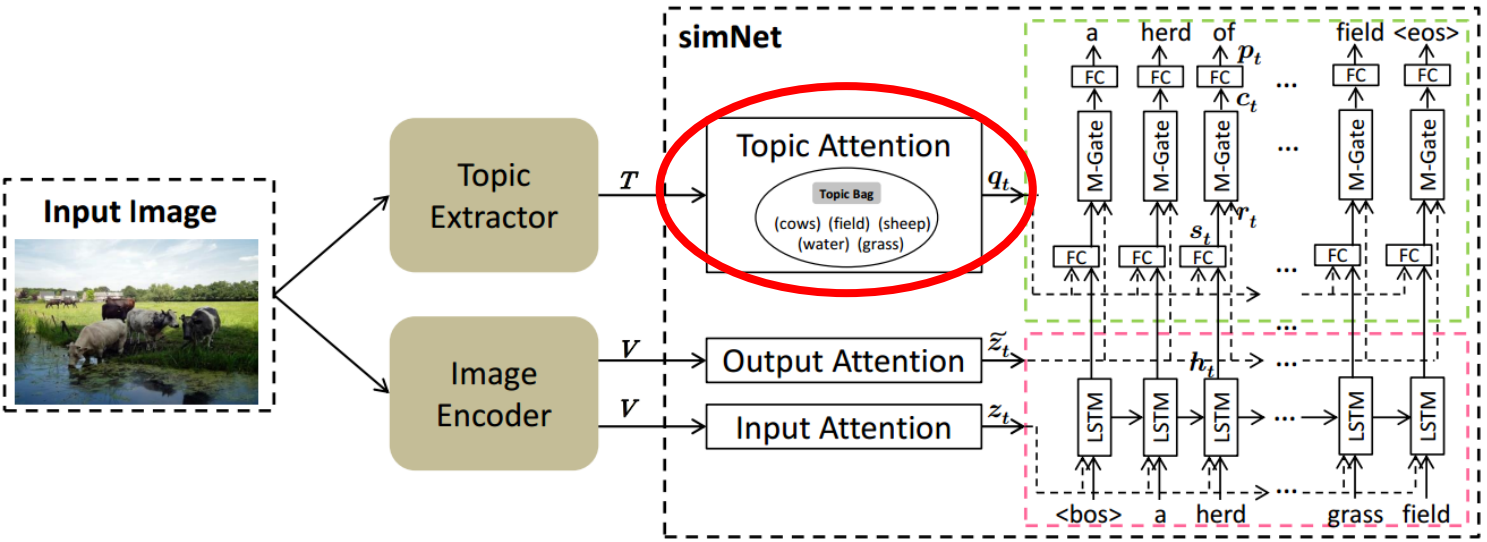
Topic Attention (Previous work):

$$\beta_t = \text{softmax}(\boldsymbol{T}^{\top} \boldsymbol{U} \boldsymbol{y}_{t-1}) \qquad (9)$$

Lacking the attentive visual information when selecting topic!

You et al., 2016 : Image captioning with semantic attention. In CVPR 2016

Approach: Our Topic Attention



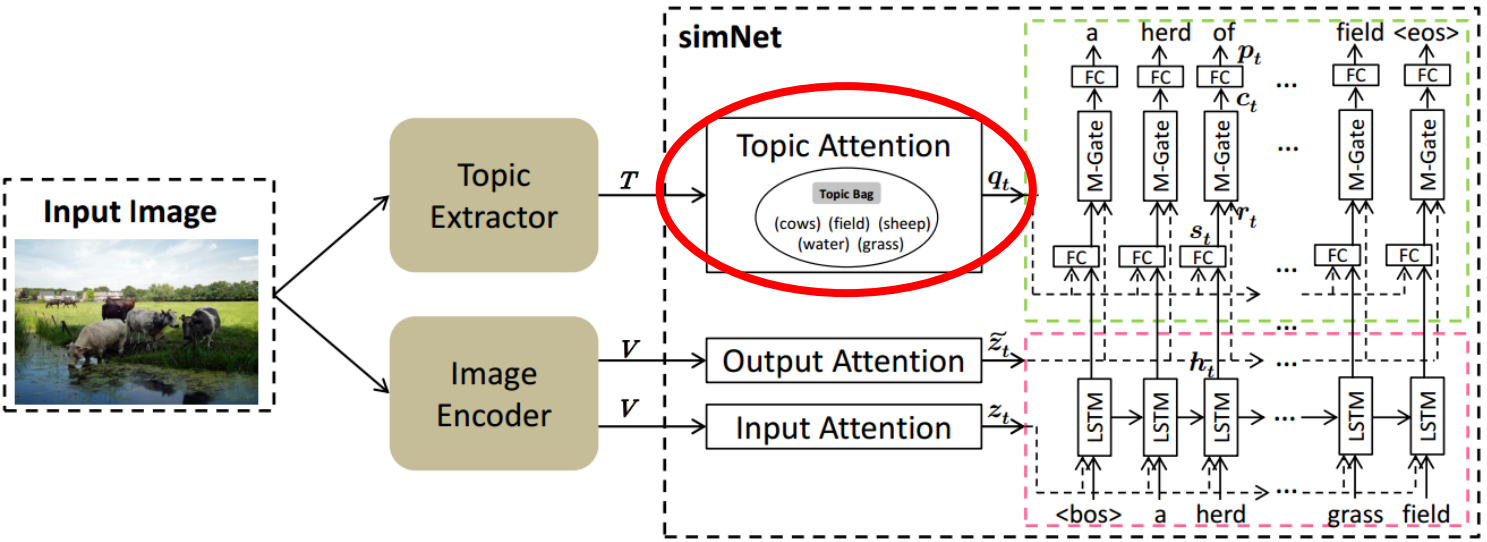
Topic Attention (Our):

$$\boldsymbol{Q}_t = \tanh(\boldsymbol{W}^{Q,T}\boldsymbol{T} \oplus \boldsymbol{W}^{Q,h}\boldsymbol{h}_t) \quad (10)$$

$$\boldsymbol{\beta}_t = \text{softmax}(\boldsymbol{Q}_t \boldsymbol{w}^{\beta,Q}) \quad (11)$$

$$\boldsymbol{q}_t = \boldsymbol{T}\boldsymbol{\beta}_t \quad (12)$$

Approach: Contextual Information



Topic Attention (Our):

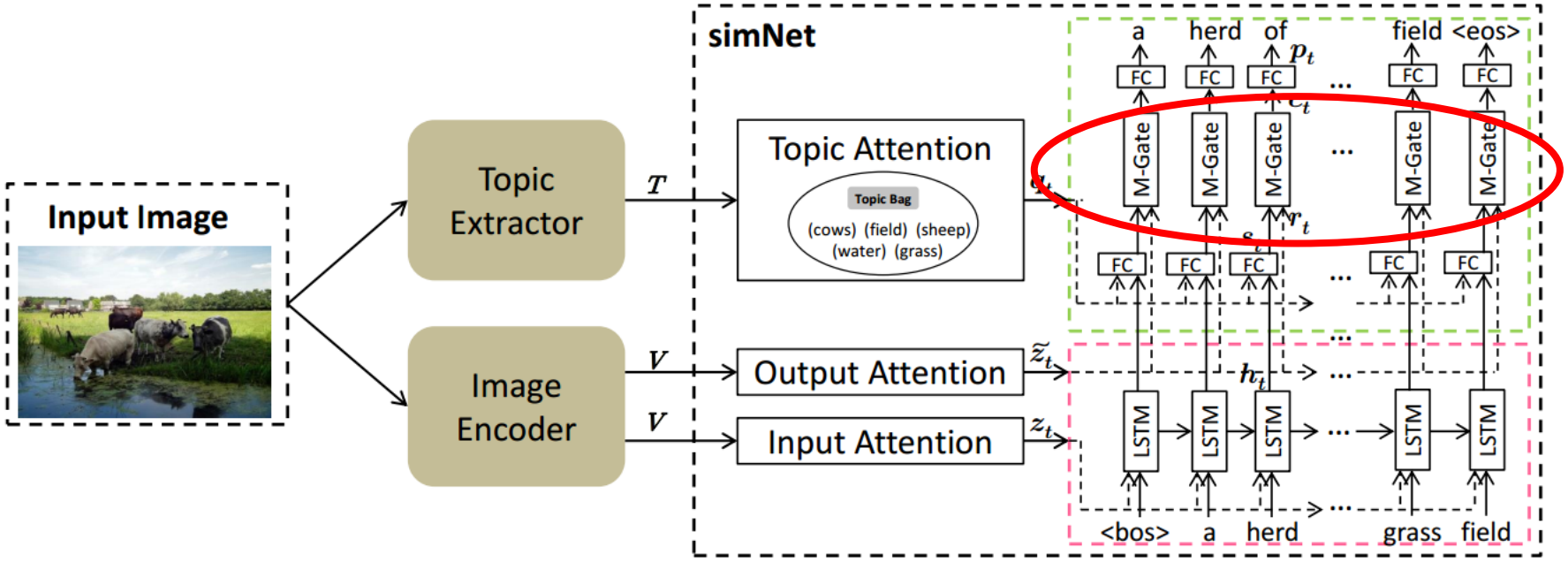
$$\boldsymbol{Q}_t = \tanh(\boldsymbol{W}^{Q,T}\boldsymbol{T} \oplus \boldsymbol{W}^{Q,h}\boldsymbol{h}_t) \quad (10)$$

$$\boldsymbol{\beta}_t = \text{softmax}(\boldsymbol{Q}_t\boldsymbol{w}^{\beta,Q}) \quad (11)$$

$$\boldsymbol{q}_t = \boldsymbol{T}\boldsymbol{\beta}_t \quad (12)$$

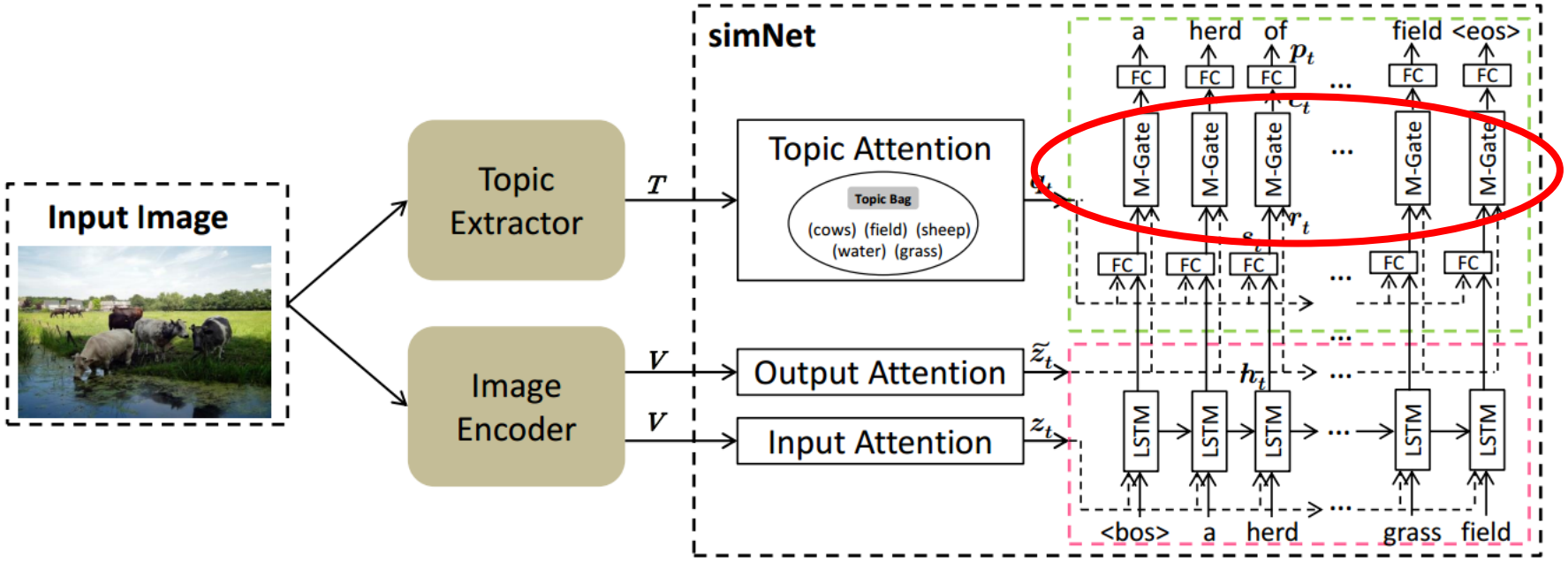
the contextual information: $\boldsymbol{s}_t = \tanh(\boldsymbol{W}^{s,q}\boldsymbol{q}_t + \boldsymbol{W}^{s,h}\boldsymbol{h}_t)$

Approach: Merging Gate



How to make full use of the visual information and the contextual information?

Approach: Merging Gate

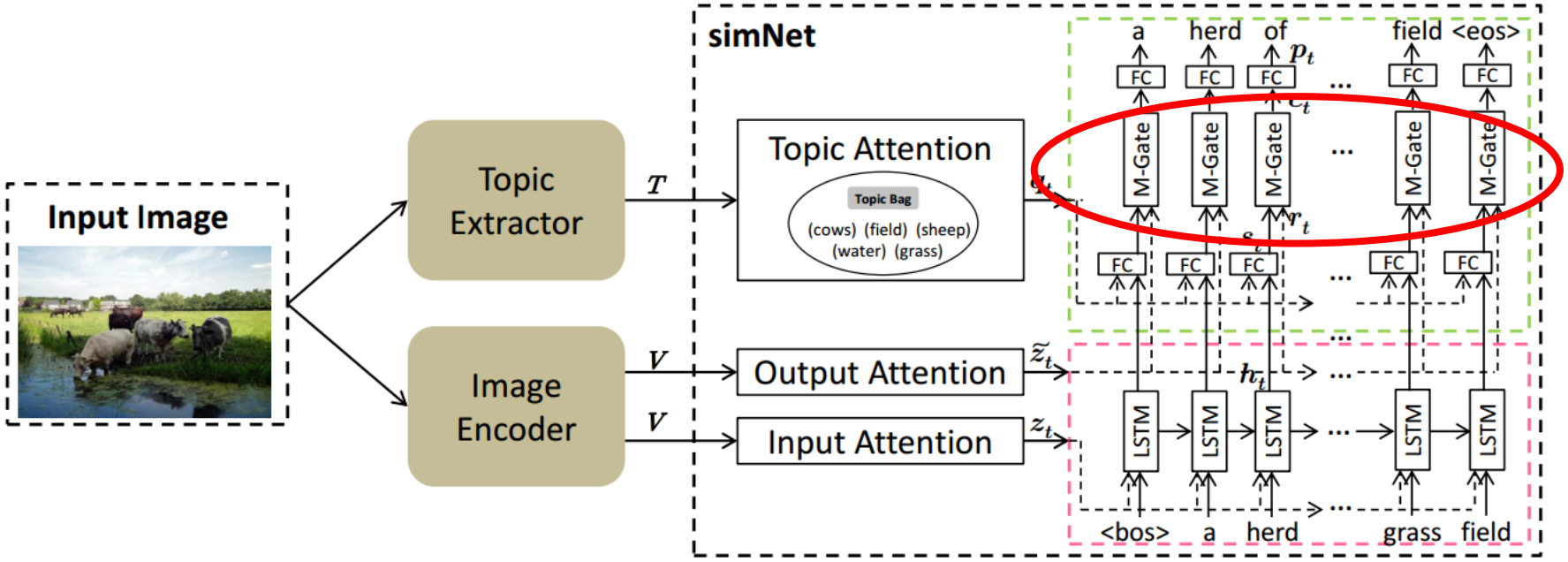


Visual information
(e.g., *"behind"*, *"red"* is better)

VS

Contextual information
(e.g., *"people"*, *"table"* is better)

Approach: Merging Gate

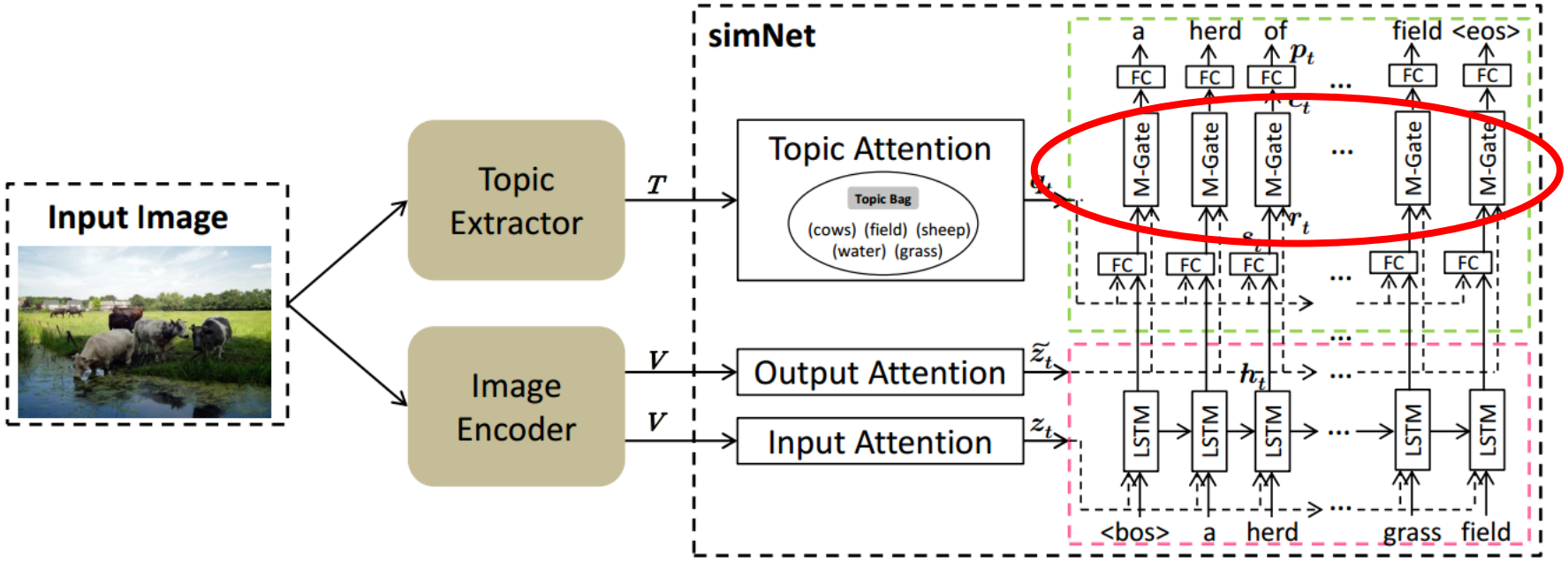


$$\gamma_t = \sigma(S(\boldsymbol{s}_t) - S(\boldsymbol{r}_t))$$

$$\boldsymbol{c}_t = \gamma_t \boldsymbol{s}_t + (1 - \gamma_t)\boldsymbol{r}_t$$

(Where σ is the sigmoid function)

Approach: Merging Gate

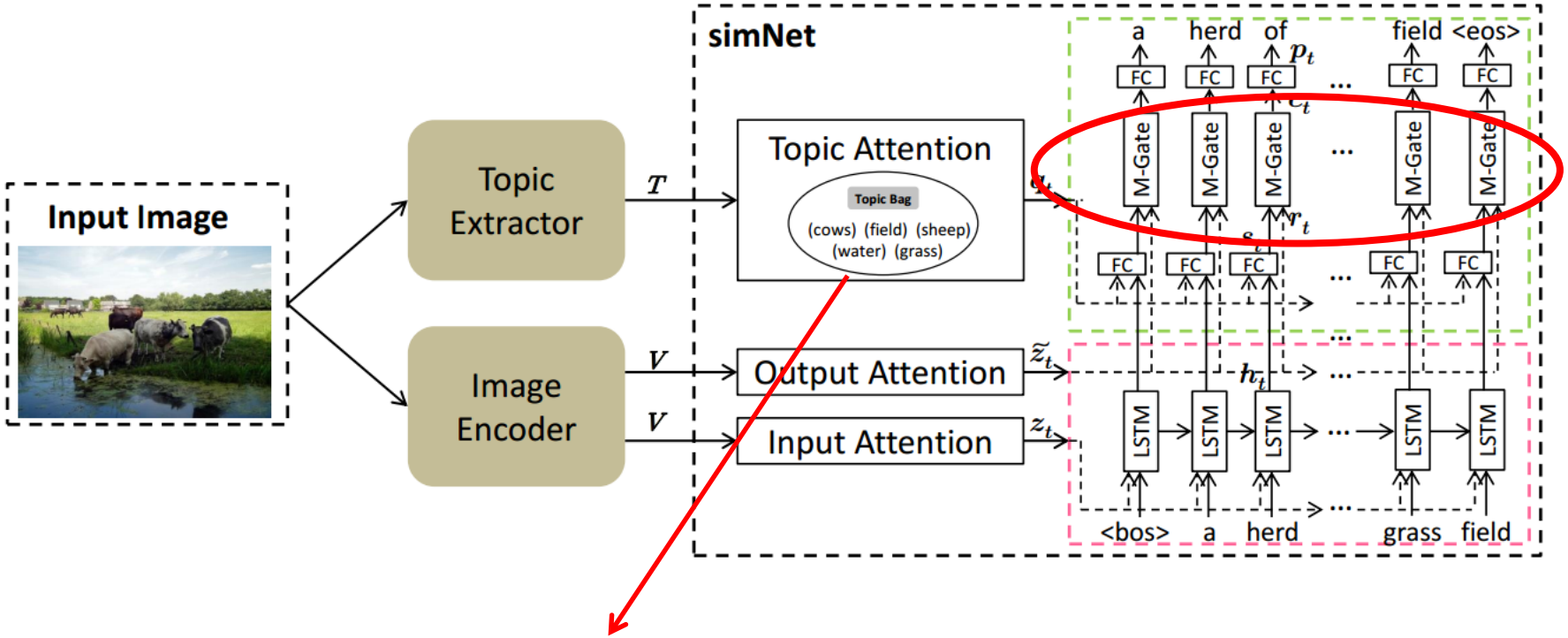


$$\gamma_t = \sigma(S(\boldsymbol{s}_t) - S(\boldsymbol{r}_t))$$

$$\boldsymbol{c}_t = \gamma_t \boldsymbol{s}_t + (1 - \gamma_t)\boldsymbol{r}_t$$

The scoring function S is designed to evaluate the importance of the topic attention.

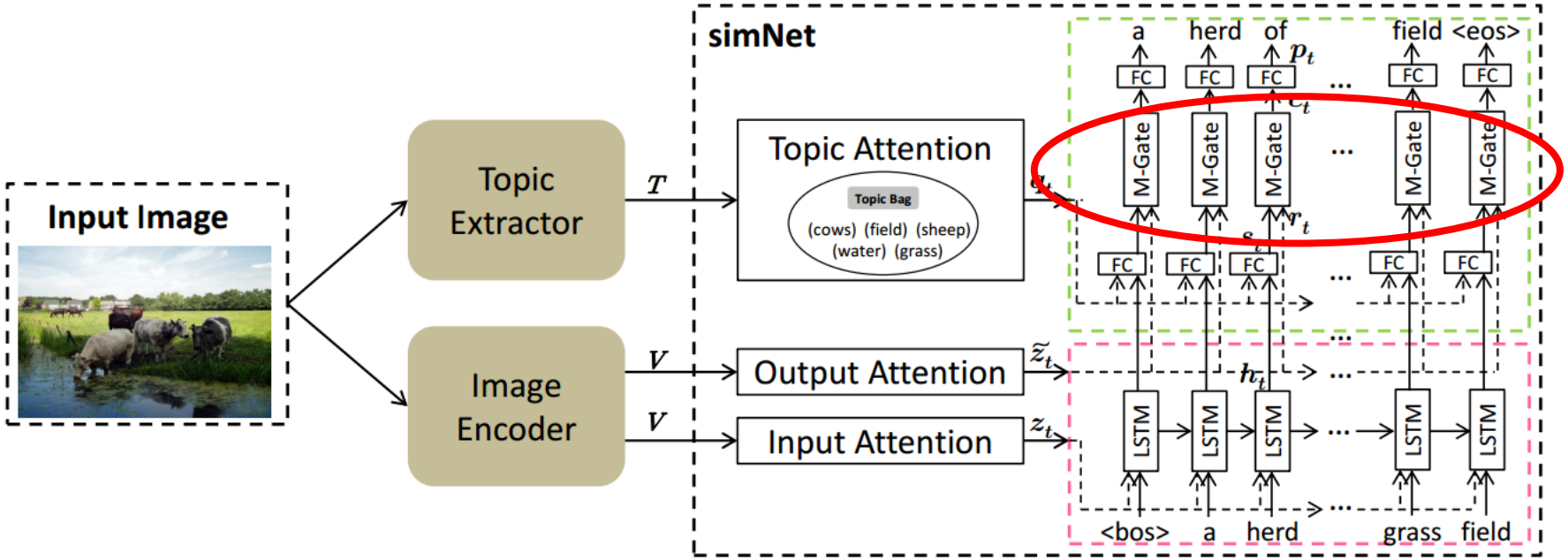
Approach: Merging Gate



$$\boldsymbol{Q}_t = \tanh(\boldsymbol{W}^{Q,T}\boldsymbol{T} \oplus \boldsymbol{W}^{Q,h}\boldsymbol{h}_t) \qquad (10)$$

$$\boldsymbol{\beta}_t = \mathrm{softmax}(\boldsymbol{Q}_t\boldsymbol{w}^{\beta,Q}) \qquad (11)$$

Approach: Merging Gate



Share Weights

$$\boldsymbol{Q}_t = \tanh(\boldsymbol{W}^{Q,T}\boldsymbol{T} \oplus \boldsymbol{W}^{Q,h}\boldsymbol{h}_t) \quad (10)$$

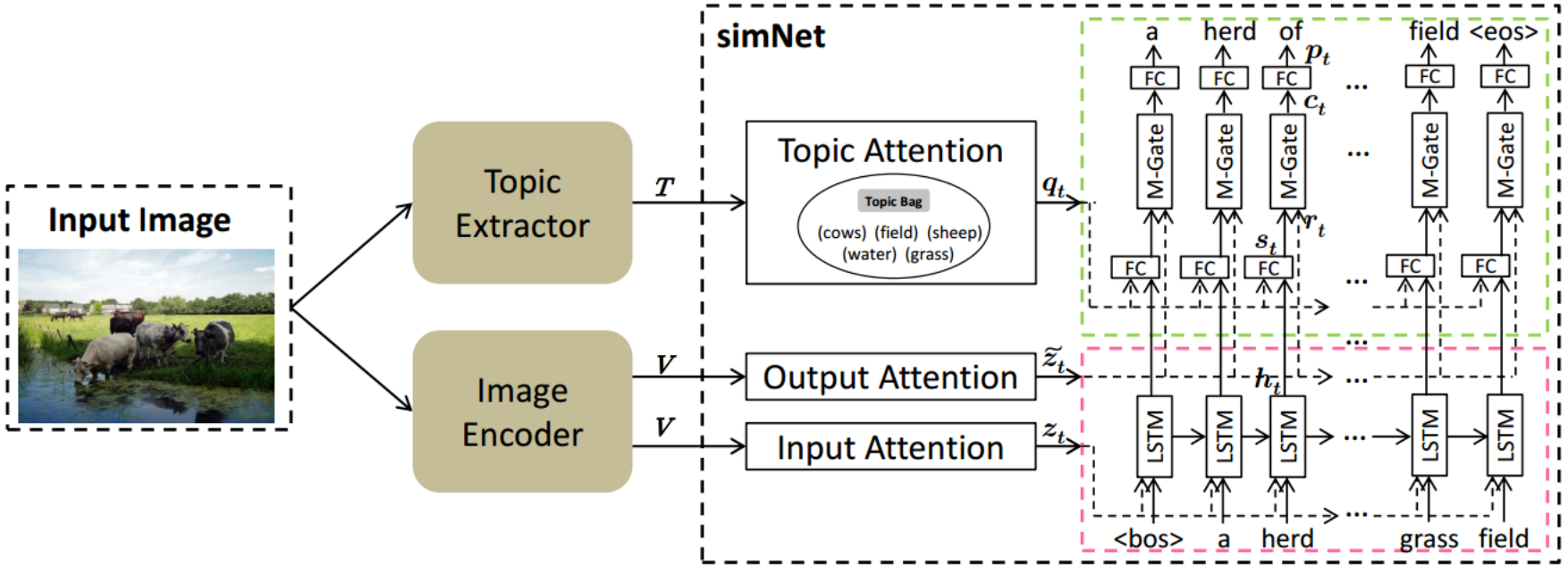
$$\boldsymbol{\beta}_t = \text{softmax}(\boldsymbol{Q}_t \boldsymbol{w}^{\beta,Q}) \quad (11)$$

$$S(\boldsymbol{s}_t) = \tanh(\boldsymbol{W}^{S,h}\boldsymbol{h}_t + \boldsymbol{W}^{S,s}\boldsymbol{s}_t) \cdot \boldsymbol{w}^S \quad (16)$$

$$S(\boldsymbol{r}_t) = \tanh(\boldsymbol{W}^{S,h}\boldsymbol{h}_t + \boldsymbol{W}^{S,r}\boldsymbol{r}_t) \cdot \boldsymbol{w}^S \quad (17)$$

Share Weights

Generating Words



the contextual information: $y_t \sim \boldsymbol{p}_t = \mathrm{softmax}(\boldsymbol{W}^{p,c}\boldsymbol{c}_t)$

3 Experiments

Experiments

Dataset

Microsoft COCO(MSCOCO) and Flickr30k

- ✓ Sparrow bird on branch, with beak inspecting leaves on branch.
- ✓ A bird sitting on the branch of a tree near leaves.
- ✓ A bird that is sitting in a tree.
- ✓ a bird sitting on a branch of a tree.
- ✓ a bird that is on a small branch of a tree.

Evaluation Metrics

- ✓ SPICE
- ✓ CIDEr
- ✓ BLEU
- ✓ METEOR
- ✓ ROUGE

Correlates the best with human Judgments !

Experiments: Results (MSCOCO)

COCO	SPICE	CIDEr	METEOR	ROUGE-L	BLEU-4
HardAtt (Xu et al., 2015)	-	-	0.230	-	0.250
ATT-FCN (You et al., 2016)	-	-	0.243	-	0.304
SCA-CNN (Chen et al., 2017)	-	0.952	0.250	0.531	0.311
LSTM-A (Yao et al., 2017)	0.186	1.002	0.254	0.540	0.326
SCN-LSTM (Gan et al., 2017)	-	1.012	0.257	-	0.330
Skeleton (Wang et al., 2017)	-	1.069	0.268	0.552	0.336
AdaAtt (Lu et al., 2017)	0.195	1.085	0.266	0.549	0.332
NBT (Lu et al., 2018)	0.201	1.072	0.271	-	0.347
DRL (Ren et al., 2017b)*	-	0.937	0.251	0.525	0.304
TD-M-ATT (Chen et al., 2018)*	-	1.116	0.268	0.555	0.336
SCST (Rennie et al., 2017)*	-	1.140	0.267	0.557	0.342
SR-PL (Liu et al., 2018)*†	0.210	1.171	0.274	**0.570**	0.358
Up-Down (Anderson et al., 2018)*†	0.214	**1.201**	0.277	0.569	**0.363**
simNet	**0.220**	1.135	**0.283**	0.564	0.332

Comparable Models

Experiments: Results (MSCOCO)

COCO	SPICE	CIDEr	METEOR	ROUGE-L	BLEU-4
HardAtt (Xu et al., 2015)	-	-	0.230	-	0.250
ATT-FCN (You et al., 2016)	-	-	0.243	-	0.304
SCA-CNN (Chen et al., 2017)	-	0.952	0.250	0.531	0.311
LSTM-A (Yao et al., 2017)	0.186	1.002	0.254	0.540	0.326
SCN-LSTM (Gan et al., 2017)	-	1.012	0.257	-	0.330
Skeleton (Wang et al., 2017)	-	1.069	0.268	0.552	0.336
AdaAtt (Lu et al., 2017)	0.195	1.085	0.266	0.549	0.332
NBT (Lu et al., 2018)	0.201	1.072	0.271	-	0.347
DRL (Ren et al., 2017b)*	-	0.937	0.251	0.525	0.304
TD-M-ATT (Chen et al., 2018)*	-	1.116	0.268	0.555	0.336
SCST (Rennie et al., 2017)*	-	1.140	0.267	0.557	0.342
SR-PL (Liu et al., 2018)*†	0.210	1.171	0.274	**0.570**	0.358
Up-Down (Anderson et al., 2018)*†	0.214	**1.201**	0.277	0.569	**0.363**
simNet	**0.220**	1.135	**0.283**	0.564	0.332

Competitive

4 Analysis

Analysis: The Contributions of The Sub-modules

Comprehensiveness

Detailedness

Methods	SPICE							CIDEr	METEOR	ROUGE-L	BLEU-4
	All	Objects	Attributes	Relations	Color	Count	Size				
Baseline (Plain Encoder-Decoder Network)	0.150	0.295	0.048	0.039	0.022	0.004	0.023	0.762	0.220	0.495	0.251
Up-Down (Anderson et al., 2018)* †	0.214	0.391	0.100	0.065	0.114	0.184	0.032	**1.201**	0.277	**0.569**	**0.363**
Baseline + Input Att.	0.164	0.316	0.060	0.044	0.030	0.038	0.024	0.840	0.233	0.512	0.273
Baseline + Output Att.	0.181	0.329	0.094	0.053	0.089	0.184	0.044	0.968	0.253	0.534	0.301
Baseline + Input Att. + Output Att.	0.187	0.338	0.101	0.055	**0.115**	0.161	**0.048**	1.038	0.259	0.542	0.311
Baseline + Topic Att.	0.184	0.348	0.074	0.051	0.047	0.064	0.037	0.915	0.250	0.517	0.260
Baseline + Topic Att. + MGate	0.189	0.355	0.080	0.051	0.055	0.090	0.033	0.959	0.256	0.527	0.281
Baseline + Input Att. + Output Att. + Topic Att.	0.206	0.381	0.091	0.060	0.075	0.094	0.045	1.068	0.273	0.556	0.320
simNet (Full Model)	**0.220**	**0.394**	**0.109**	**0.070**	0.088	**0.202**	0.045	1.135	**0.283**	0.564	0.332

Analysis: Output Attention

The output attention is much more effective than the input attention

Methods	SPICE							CIDEr	METEOR	ROUGE-L	BLEU-4
	All	Objects	Attributes	Relations	Color	Count	Size				
Baseline (Plain Encoder-Decoder Network)	0.150	0.295	0.048	0.039	0.022	0.004	0.023	0.762	0.220	0.495	0.251
Up-Down (Anderson et al., 2018)* †	0.214	0.391	0.100	0.065	0.114	0.184	0.032	**1.201**	0.277	**0.569**	**0.363**
Baseline + Input Att.	0.164	0.316	0.060	0.044	0.030	0.038	0.024	0.840	0.233	0.512	0.273
Baseline + Output Att.	0.181	0.329	0.094	0.053	0.089	0.184	0.044	0.968	0.253	0.534	0.301
Baseline + Input Att. + Output Att.	0.187	0.338	0.101	0.055	**0.115**	0.161	**0.048**	1.038	0.259	0.542	0.311
Baseline + Topic Att.	0.184	0.348	0.074	0.051	0.047	0.064	0.037	0.915	0.250	0.517	0.260
Baseline + Topic Att. + MGate	0.189	0.355	0.080	0.051	0.055	0.090	0.033	0.959	0.256	0.527	0.281
Baseline + Input Att. + Output Att. + Topic Att.	0.206	0.381	0.091	0.060	0.075	0.094	0.045	1.068	0.273	0.556	0.320
simNet (Full Model)	**0.220**	**0.394**	**0.109**	**0.070**	0.088	**0.202**	0.045	1.135	**0.283**	0.564	0.332

Analysis: Visual Attention

A combination of the input attention and the output attention makes the results even better

Methods	SPICE							CIDEr	METEOR	ROUGE-L	BLEU-4
	All	Objects	Attributes	Relations	Color	Count	Size				
Baseline (Plain Encoder-Decoder Network)	0.150	0.295	0.048	0.039	0.022	0.004	0.023	0.762	0.220	0.495	0.251
Up-Down (Anderson et al., 2018)* †	0.214	0.391	0.100	0.065	0.114	0.184	0.032	**1.201**	0.277	**0.569**	**0.363**
Baseline + Input Att.	0.164	0.316	0.060	0.044	0.030	0.038	0.024	0.840	0.233	0.512	0.273
Baseline + Output Att.	0.181	0.329	0.094	0.053	0.089	0.184	0.044	0.968	0.253	0.534	0.301
Baseline + Input Att. + Output Att.	0.187	0.338	0.101	0.055	**0.115**	0.161	**0.048**	1.038	0.259	0.542	0.311
Baseline + Topic Att.	0.184	0.348	0.074	0.051	0.047	0.064	0.037	0.915	0.250	0.517	0.260
Baseline + Topic Att. + MGate	0.189	0.355	0.080	0.051	0.055	0.090	0.033	0.959	0.256	0.527	0.281
Baseline + Input Att. + Output Att. + Topic Att.	0.206	0.381	0.091	0.060	0.075	0.094	0.045	1.068	0.273	0.556	0.320
simNet (Full Model)	**0.220**	**0.394**	**0.109**	**0.070**	0.088	**0.202**	0.045	1.135	**0.283**	0.564	0.332

Analysis: Topic Attention

The topic attention is better at identifying objects but worse at identifying attributes.

Methods	SPICE							CIDEr	METEOR	ROUGE-L	BLEU-4
	All	Objects	Attributes	Relations	Color	Count	Size				
Baseline (Plain Encoder-Decoder Network)	0.150	0.295	0.048	0.039	0.022	0.004	0.023	0.762	0.220	0.495	0.251
Up-Down (Anderson et al., 2018)* †	0.214	0.391	0.100	0.065	0.114	0.184	0.032	**1.201**	0.277	**0.569**	**0.363**
Baseline + Input Att.	0.164	0.316	0.060	0.044	0.030	0.038	0.024	0.840	0.233	0.512	0.273
Baseline + Output Att.	0.181	0.329	0.094	0.053	0.089	0.184	0.044	0.968	0.253	0.534	0.301
Baseline + Input Att. + Output Att.	0.187	0.338	0.101	0.055	**0.115**	0.161	**0.048**	1.038	0.259	0.542	0.311
Baseline + Topic Att.	0.184	0.348	0.074	0.051	0.047	0.064	0.037	0.915	0.250	0.517	0.260
Baseline + Topic Att. + MGate	0.189	0.355	0.080	0.051	0.055	0.090	0.033	0.959	0.256	0.527	0.281
Baseline + Input Att. + Output Att. + Topic Att.	0.206	0.381	0.091	0.060	0.075	0.094	0.045	1.068	0.273	0.556	0.320
simNet (Full Model)	**0.220**	**0.394**	**0.109**	**0.070**	0.088	**0.202**	0.045	1.135	**0.283**	0.564	0.332

Analysis: Visual Attention + Topic Attention

Combing the visual attention and the topic attention directly results in a huge boost in performance

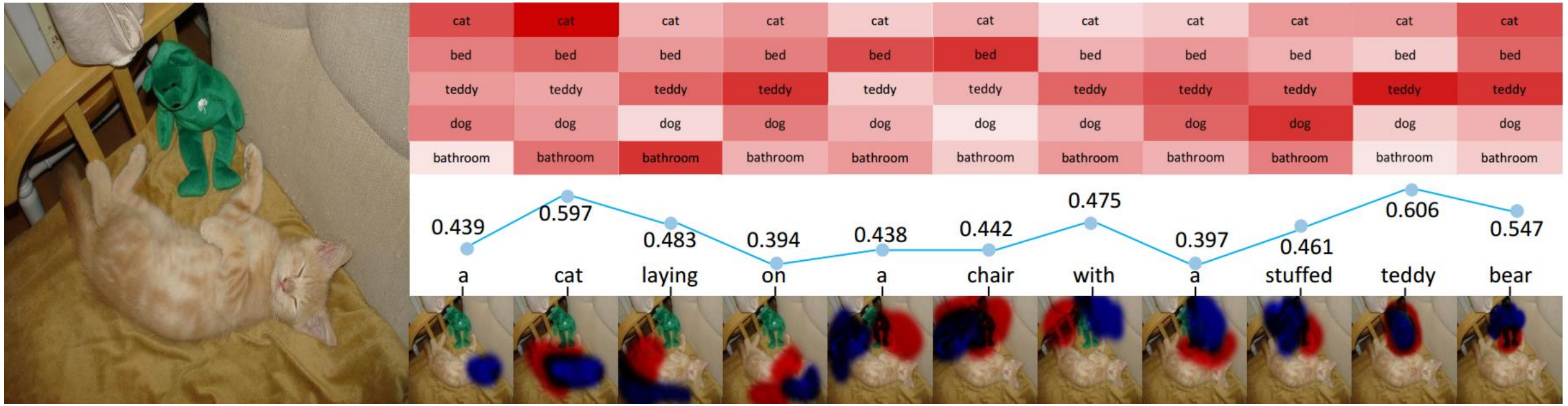
Methods	SPICE							CIDEr	METEOR	ROUGE-L	BLEU-4
	All	Objects	Attributes	Relations	Color	Count	Size				
Baseline (Plain Encoder-Decoder Network)	0.150	0.295	0.048	0.039	0.022	0.004	0.023	0.762	0.220	0.495	0.251
Up-Down (Anderson et al., 2018)* †	0.214	0.391	0.100	0.065	0.114	0.184	0.032	**1.201**	0.277	**0.569**	**0.363**
Baseline + Input Att.	0.164	0.316	0.060	0.044	0.030	0.038	0.024	0.840	0.233	0.512	0.273
Baseline + Output Att.	0.181	0.329	0.094	0.053	0.089	0.184	0.044	0.968	0.253	0.534	0.301
Baseline + Input Att. + Output Att.	0.187	0.338	0.101	0.055	**0.115**	0.161	**0.048**	1.038	0.259	0.542	0.311
Baseline + Topic Att.	0.184	0.348	0.074	0.051	0.047	0.064	0.037	0.915	0.250	0.517	0.260
Baseline + Topic Att. + MGate	0.189	0.355	0.080	0.051	0.055	0.090	0.033	0.959	0.256	0.527	0.281
Baseline + Input Att. + Output Att. + Topic Att.	0.206	0.381	0.091	0.060	0.075	0.094	0.045	1.068	0.273	0.556	0.320
simNet (Full Model)	**0.220**	**0.394**	**0.109**	**0.070**	0.088	**0.202**	0.045	1.135	**0.283**	0.564	0.332

Analysis: Full Model

Applying the merging gate is essential to the overall performance.

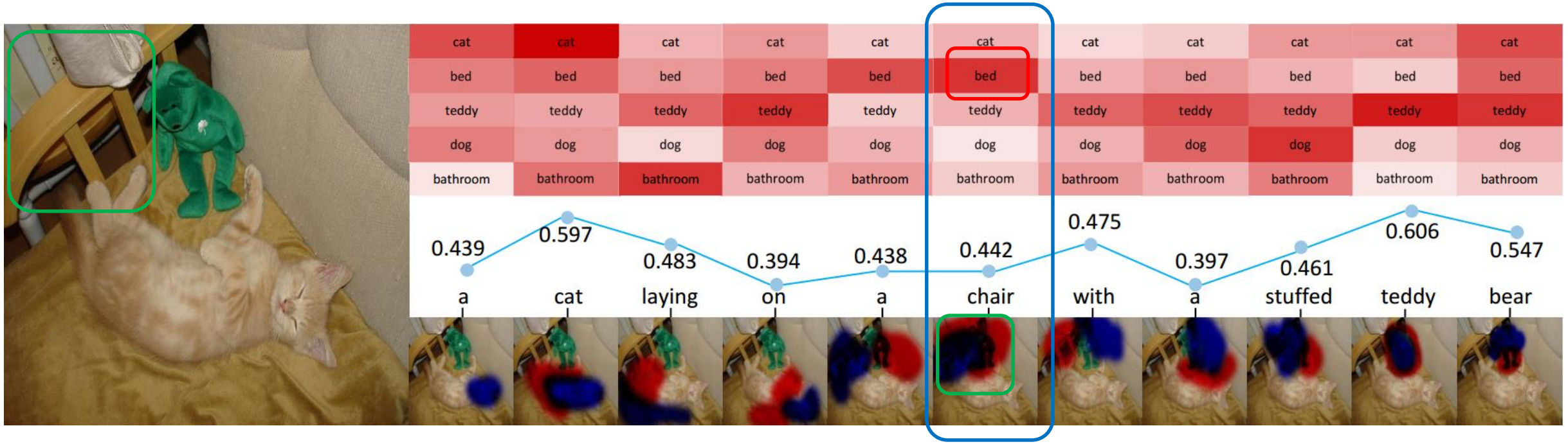
Methods	SPICE							CIDEr	METEOR	ROUGE-L	BLEU-4
	All	Objects	Attributes	Relations	Color	Count	Size				
Baseline (Plain Encoder-Decoder Network)	0.150	0.295	0.048	0.039	0.022	0.004	0.023	0.762	0.220	0.495	0.251
Up-Down (Anderson et al., 2018)* †	0.214	0.391	0.100	0.065	0.114	0.184	0.032	**1.201**	0.277	**0.569**	**0.363**
Baseline + Input Att.	0.164	0.316	0.060	0.044	0.030	0.038	0.024	0.840	0.233	0.512	0.273
Baseline + Output Att.	0.181	0.329	0.094	0.053	0.089	0.184	0.044	0.968	0.253	0.534	0.301
Baseline + Input Att. + Output Att.	0.187	0.338	0.101	0.055	**0.115**	0.161	**0.048**	1.038	0.259	0.542	0.311
Baseline + Topic Att.	0.184	0.348	0.074	0.051	0.047	0.064	0.037	0.915	0.250	0.517	0.260
Baseline + Topic Att. + MGate	0.189	0.355	0.080	0.051	0.055	0.090	0.033	0.959	0.256	0.527	0.281
Baseline + Input Att. + Output Att. + Topic Att.	0.206	0.381	0.091	0.060	0.075	0.094	0.045	1.068	0.273	0.556	0.320
simNet (Full Model)	**0.220**	**0.394**	**0.109**	**0.070**	0.088	**0.202**	0.045	1.135	**0.283**	0.564	0.332

Analysis: Visualization



- The upper part shows the attention weights of each of 5 extracted topics. → Deeper color means larger in value.
- The middle part shows the value of the merging gate. → Determines the importance of the topic attention.
- The lower part shows the visualization of visual attention. → The blue shade indicates the output attention. The red shade indicates the input attention.

Analysis: Visualization



Visual information "*chair*" is more important than contextual information "*bed*"

Analysis: Examples

Comparison of Models			
Topics	woman girl baby bear kitchen	computer keyboard laptop mouse desk	pizza cheese table plate toppings
Visual Attention	a girl and a baby are holding a stuffed animal	a computer keyboard sitting on top of a wooden desk → lacking "mouse"	two pizzas with toppings on a table
Topic Attention	a woman holding a teddy bear in a kitchen → erroneous topic "kitchen"	a computer keyboard and a mouse sitting on a desk → missing "wooden"	a pizza with a lot of toppings on it → error count
simNet	a woman and a baby are holding a stuffed animal	a computer keyboard and mouse on a wooden desk	two pizzas sitting on a table with two different kinds of toppings

Conclusion

- Stepwise image-topic merging network can adaptively combine the visual and the semantic attention to achieve substantial improvements.
- The generated captions are both detailed and comprehensive
- Our approach outperforms previous works in terms of SPICE on COCO and Flickr datasets.

Thank you!

If you have any questions about our paper, you can send a email to lfl@bupt.edu.cn

Experiments: Results (Flickr30k)

Flickr30k	SPICE	CIDEr	METEOR	ROUGE-L	BLEU-4
HardAtt (Xu et al., 2015)	-	-	0.185	-	0.199
SCA-CNN (Chen et al., 2017)	-	-	0.195	-	0.223
ATT-FCN (You et al., 2016)	-	-	0.189	-	0.230
SCN-LSTM (Gan et al., 2017)	-	-	0.210	-	0.257
AdaAtt (Lu et al., 2017)	0.145	0.531	0.204	0.467	0.251
NBT (Lu et al., 2018)	0.156	0.575	0.217	-	0.271
SR-PL (Liu et al., 2018)*†	0.158	**0.650**	0.218	**0.499**	**0.293**
simNet	**0.160**	0.585	**0.221**	0.489	0.251

Table 1: Performance on the Flickr30k Karpathy test split. The symbol * denotes directly optimizing CIDEr. The symbol † denotes using extra data for training, thus not directly comparable. Nonetheless, our model supersedes all existing models in SPICE, which correlates the best with human judgments.

Analysis: Topic Extraction

Method	Precision	Recall	F1
Topics (m=5)	49.95	38.91	42.48
All words (m=5)	**84.01**	17.99	29.49
All words (m=10)	70.90	30.18	42.05
All words (m=20)	52.51	**44.53**	**47.80**

Table 4: Performance of visual word extraction.

Method	S	C	M	R	B
Topics (m=5)	**0.220**	**1.135**	**0.283**	**0.564**	**0.332**
All words (m=5)	0.197	1.047	0.264	0.550	0.314
All words (m=10)	0.201	1.076	0.256	0.528	0.293
All words (m=20)	0.209	1.117	0.276	0.561	0.329

Table 5: Effect of using different visual words.

The reason of using objects as topics is that they are easier to identify so that the generation suffers less from erroneous predictions.

It proves that for semantic attention, it is also important to limit the absolute number of incorrect visual words instead of merely the precision or the recall.

Analysis: Merging Gate

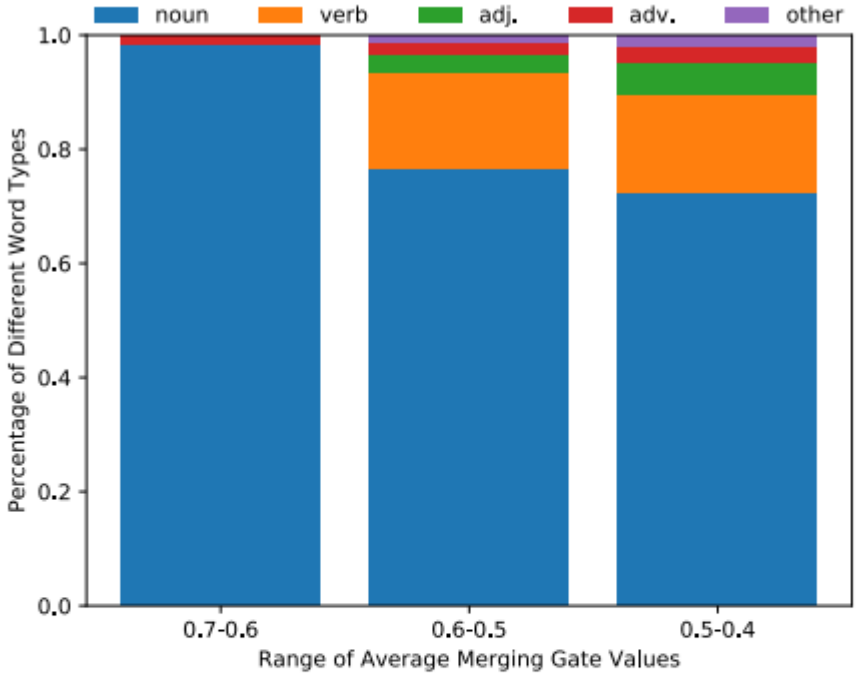
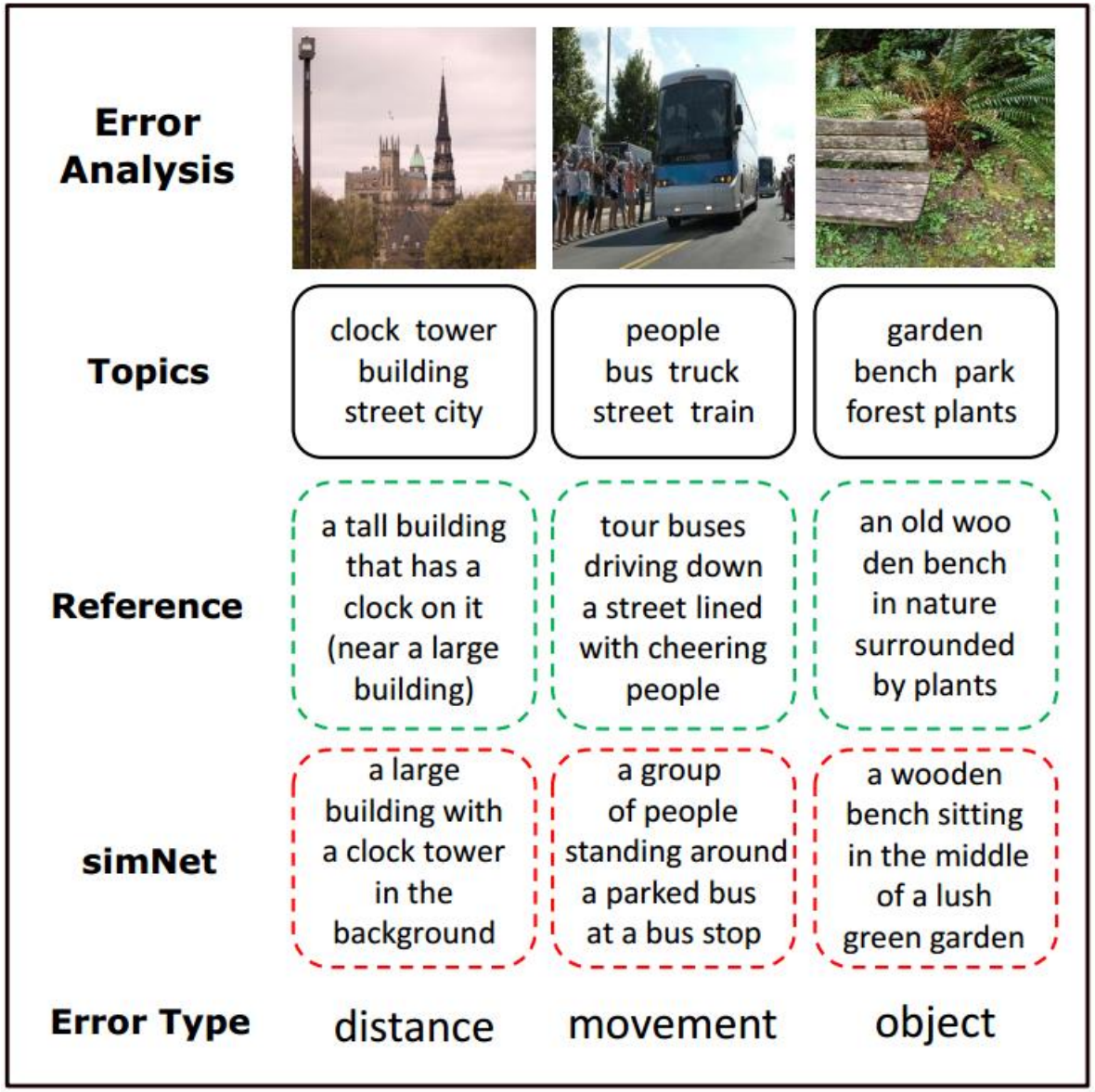


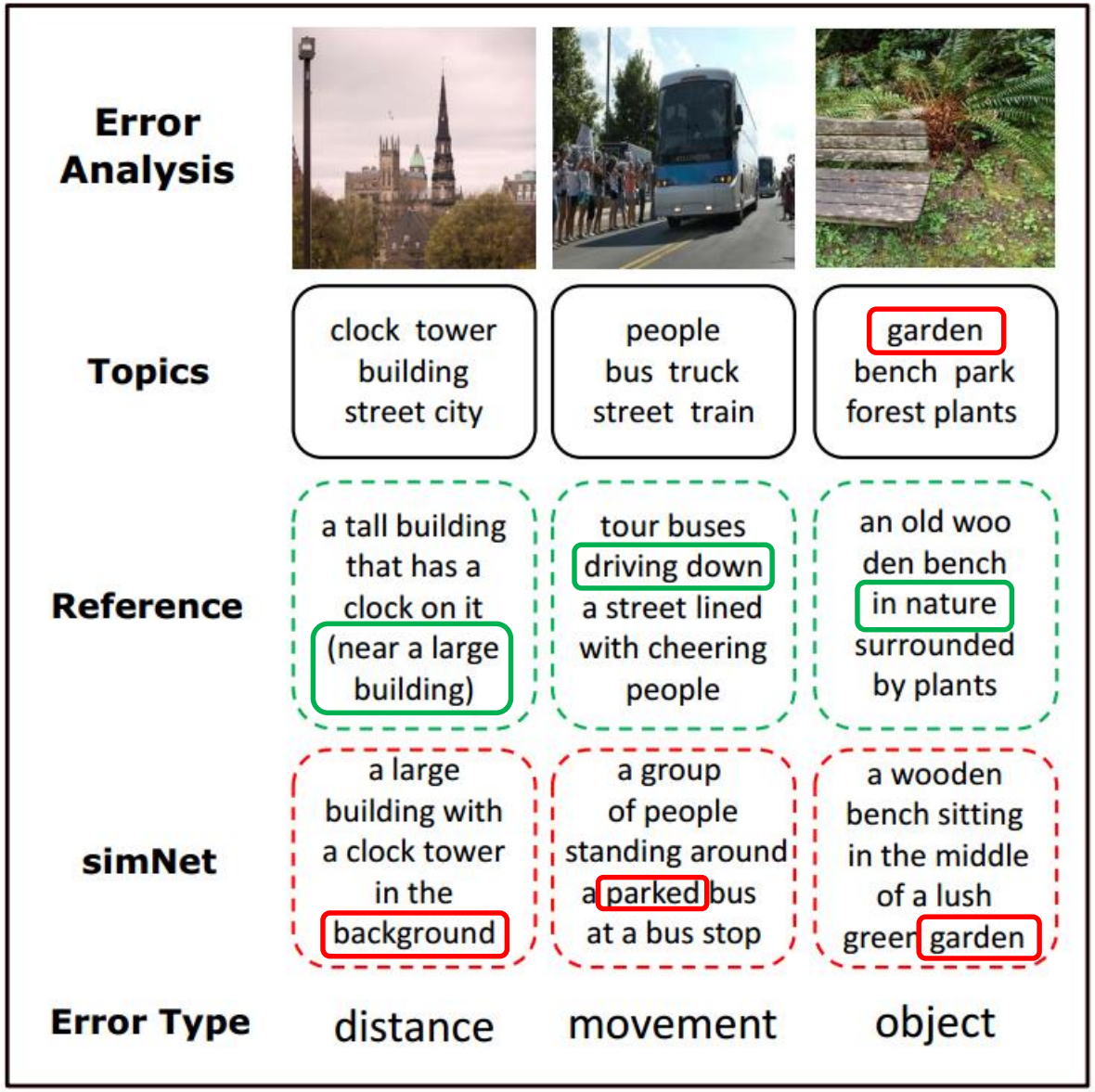
Figure 4: Average merging gate values according to word types. As we can see, object words (noun) dominate the high value range, while attribute and relation words are assigned lower values, indicating the merging gate learns to efficiently combine the information.

Analysis: Error Analysis



There are mainly three types of errors, i.e. distance (32, 26%), movement (22,18%), and object (60, 49%), with 9 (7%) other errors.

Analysis: Error Analysis



There are mainly three types of errors, i.e. distance (32, 26%), movement (22,18%), and object (60, 49%), with 9 (7%) other errors.

Experiments: Results (MSCOCO)

COCO	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
HardAtt (Xu et al., 2015)	0.705	0.881	0.528	0.779	0.383	0.658	0.277	0.537	0.241	0.322	0.516	0.654	0.865	0.893
ATT-FCN (You et al., 2016)	0.731	0.900	0.565	0.815	0.424	0.709	0.316	0.599	0.250	0.335	0.535	0.682	0.943	0.958
SCA-CNN (Chen et al., 2017)	0.712	0.894	0.542	0.802	0.404	0.691	0.302	0.579	0.244	0.331	0.524	0.674	0.912	0.921
LSTM-A (Yao et al., 2017)	0.739	0.919	0.575	0.842	0.436	0.740	0.330	0.632	0.256	0.350	0.542	0.700	0.984	1.003
SCN-LSTM (Gan et al., 2017)	0.740	0.917	0.575	0.839	0.436	0.739	0.331	0.631	0.257	0.348	0.543	0.696	1.003	1.013
AdaAtt (Lu et al., 2017)†	0.748	0.920	0.584	0.845	0.444	0.744	0.336	0.637	0.264	0.359	0.550	0.705	1.042	1.059
TD-M-ATT (Chen et al., 2018)*†	0.757	0.913	0.591	0.836	0.441	0.726	0.324	0.609	0.259	0.342	0.547	0.689	1.059	1.090
SCST (Rennie et al., 2017)*†	0.781	0.937	0.619	0.860	0.470	0.759	0.352	0.645	0.270	0.355	0.563	0.707	1.147	1.167
Up-Down (Anderson et al., 2018)*†‡	**0.802**	**0.952**	**0.641**	**0.888**	**0.491**	**0.794**	**0.369**	**0.685**	**0.276**	**0.367**	**0.571**	**0.724**	**1.179**	**1.205**
simNet	0.766	0.941	0.605	0.874	0.462	0.778	0.350	0.671	0.267	0.362	0.558	0.716	1.087	1.111

Table 6: Performance on the online COCO evaluation server. The SPICE metric is unavailable for our model, thus not reported. c5 means evaluating against 5 references, and c40 means evaluating against 40 references. The symbol * denotes directly optimizing CIDEr. The symbol † denotes model ensemble. The symbol ‡ denotes using extra data for training, thus not directly comparable. Our submission does not use the three aforementioned techniques. Nonetheless, our model is second only to Up-Down and surpasses almost all the other models in published work, especially when 40 references are considered.

Analysis: Topic Attention

Methods	SPICE							CIDEr	METEOR	ROUGE-L	BLEU-4
	All	Objects	Attributes	Relations	Color	Count	Size				
Baseline (Plain Encoder-Decoder Network)	0.150	0.295	0.048	0.039	0.022	0.004	0.023	0.762	0.220	0.495	0.251
Up-Down (Anderson et al., 2018)* †	0.214	0.391	0.100	0.065	0.114	0.184	0.032	**1.201**	0.277	**0.569**	**0.363**
Baseline + Input Att.	0.164	0.316	0.060	0.044	0.030	0.038	0.024	0.840	0.233	0.512	0.273
Baseline + Output Att.	0.181	0.329	0.094	0.053	0.089	0.184	0.044	0.968	0.253	0.534	0.301
Baseline + Input Att. + Output Att.	0.187	0.338	0.101	0.055	**0.115**	0.161	**0.048**	1.038	0.259	0.542	0.311
Baseline + Topic Att.	0.184	0.348	0.074	0.051	0.047	0.064	0.037	0.915	0.250	0.517	0.260
Baseline + Topic Att. + MGate	0.189	0.355	0.080	0.051	0.055	0.090	0.033	0.959	0.256	0.527	0.281
Baseline + Input Att. + Output Att. + Topic Att.	0.206	0.381	0.091	0.060	0.075	0.094	0.045	1.068	0.273	0.556	0.320
simNet (Full Model)	**0.220**	**0.394**	**0.109**	**0.070**	0.088	**0.202**	0.045	1.135	**0.283**	0.564	0.332