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

Search

컴퓨터과학과 202132033 염지현

Liu, Fenglin, et al. "simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions." arXiv preprint arXiv:1808.08732 (2018).

Contents

Search

Contents

Search

01 Introduction (1) 하고자 하는 것

Search

- Image captioning: 이미지를 설명하는 text 생성 프로세스로 NLP와 CV의 결합

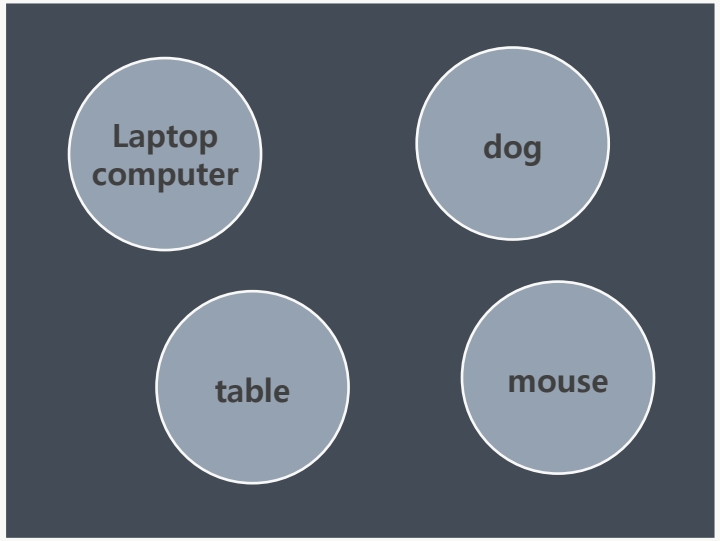
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

01 Introduction (1) 하고자 하는 것

Input image
(시각 정보)



Topic
(의미 정보)

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

Output

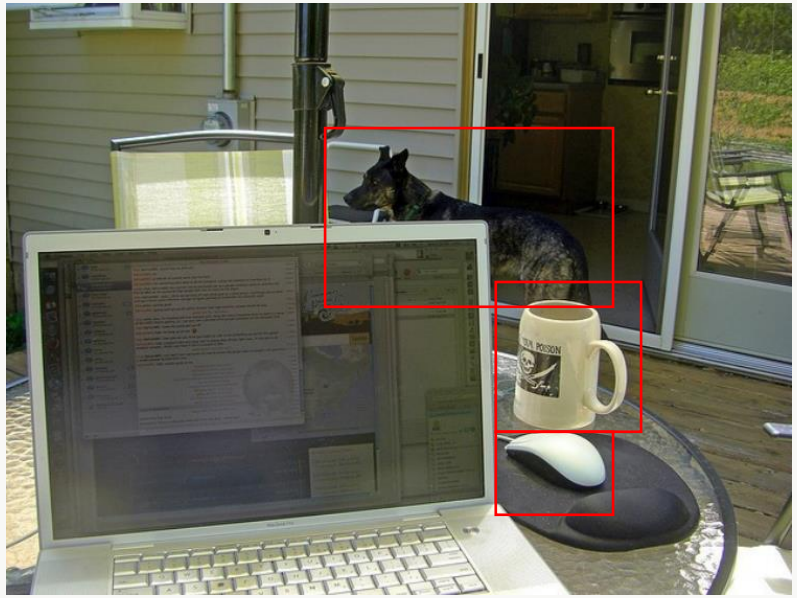
01 Introduction (2) 기존 연구의 한계점

Search

method	model	output	detail
Soft-Attention [Xu et al., 2015]	**시각적 정보** 기반 모델	a open laptop computer sitting on top of a table	• 비교적 자연스러운 문장 • mouse, cup 등 요소 생략
ATT-FCN [You et al., 2016]	**개념 정보** 기반 모델	a dog sitting on a desk with a laptop computer and mouse	• 비교적 많은 요소 포함 • 순서 미지정으로 인해 개체와 세부 사항 연결성의 한계

01 Introduction (2) 기존 연구의 한계점

Search



method	model	output	detail
Soft-Attention [Xu et al., 2015]	**시각적 정보** 기반 모델	a open laptop computer sitting on top of a table	• 비교적 자연스러운 문장 • mouse, cup 등 요소 생략
ATT-FCN [You et al., 2016]	**개념 정보** 기반 모델	a dog sitting on a desk with a laptop computer and mouse	• 비교적 많은 요소 포함 • 순서 미지정으로 인해 개체와 세부 사항 연결성의 한계

01 Introduction (2) 기존 연구의 한계점

Search

method	model	output	detail
Soft-Attention [Xu et al., 2015]	**시각적 정보** 기반 모델	a open laptop computer sitting on top of a table	• 비교적 자연스러운 문장 • mouse, cup 등 요소 생략
ATT-FCN [You et al., 2016]	**개념 정보** 기반 모델	a dog sitting on a desk with a laptop computer and mouse	• 비교적 많은 요소 포함 • 순서 미지정으로 인해 개체와 세부 사항 연결성의 한계

01 Introduction (3) 해결 방법

Search

시각 정보

: 객체의 속성과 관계에 초점

+

의미 정보

: 포괄적인 정보 제공

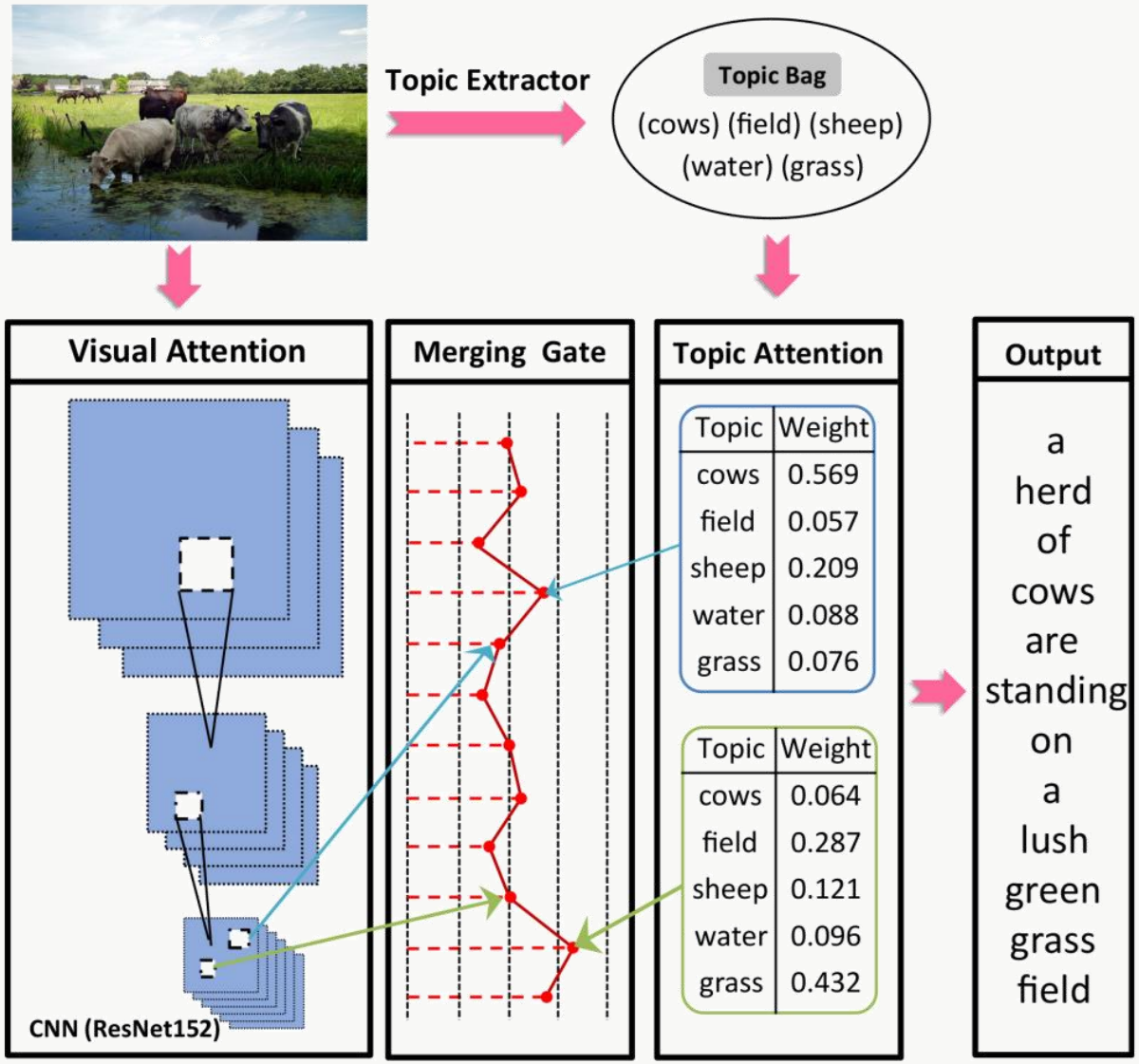
상세하고 포괄적인 image captions 생성

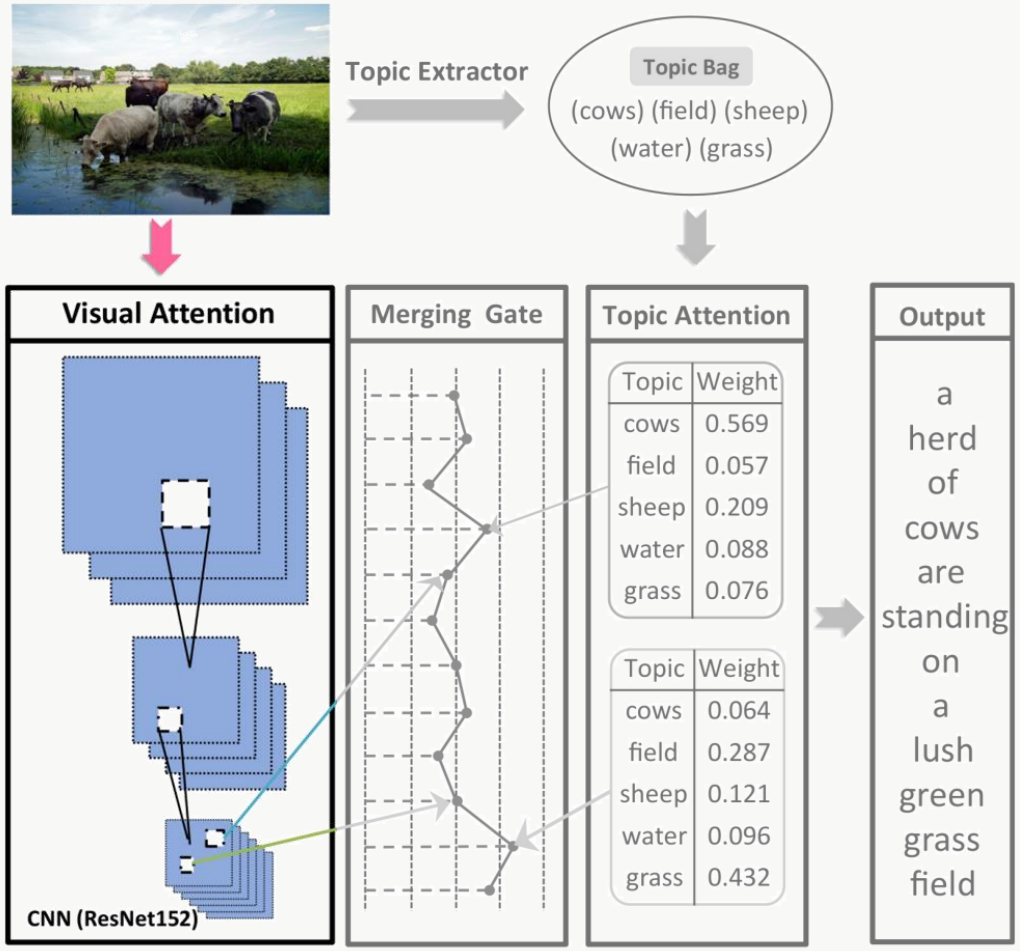
Contents

Search

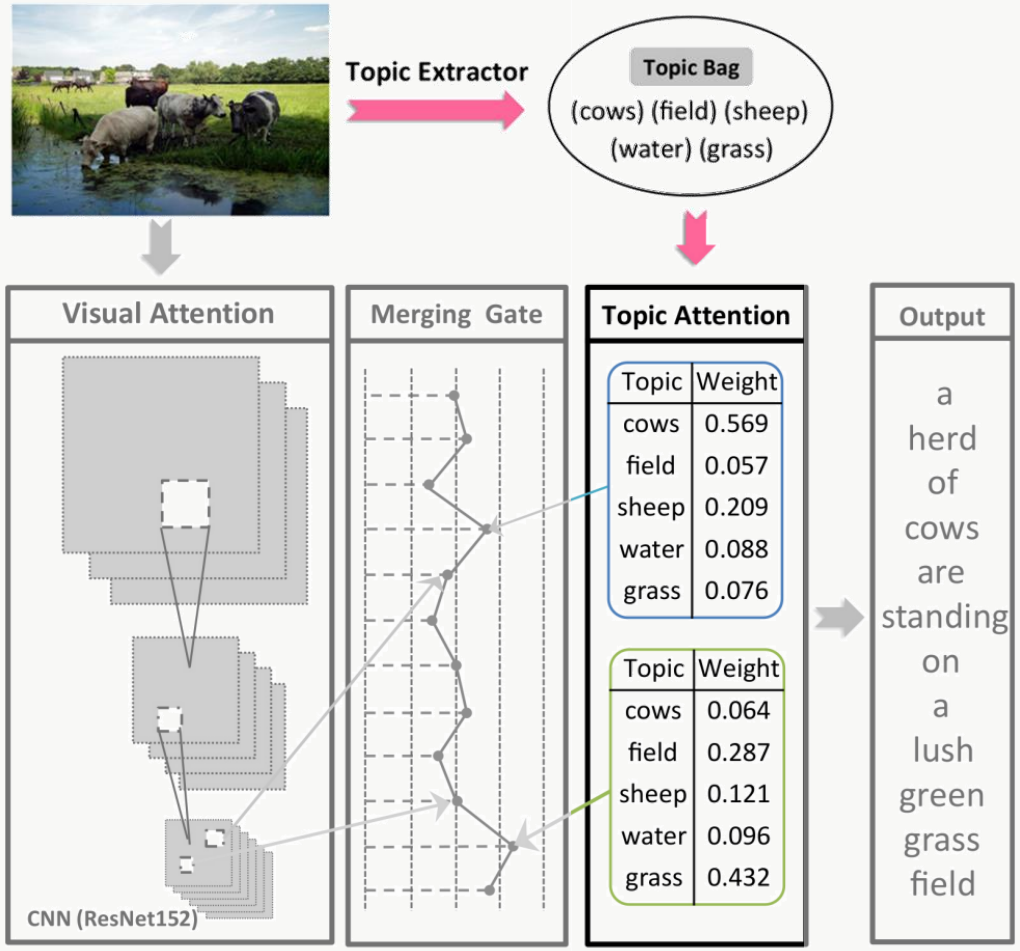
02 Method

Search

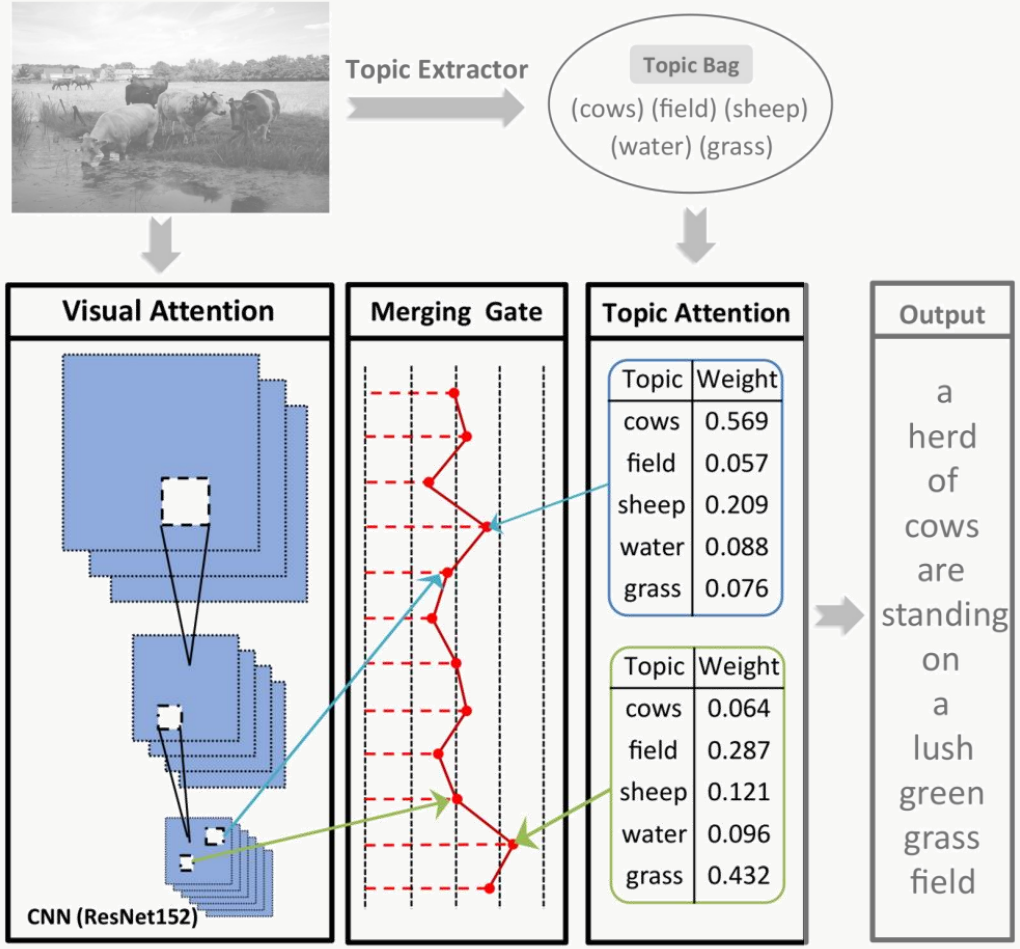




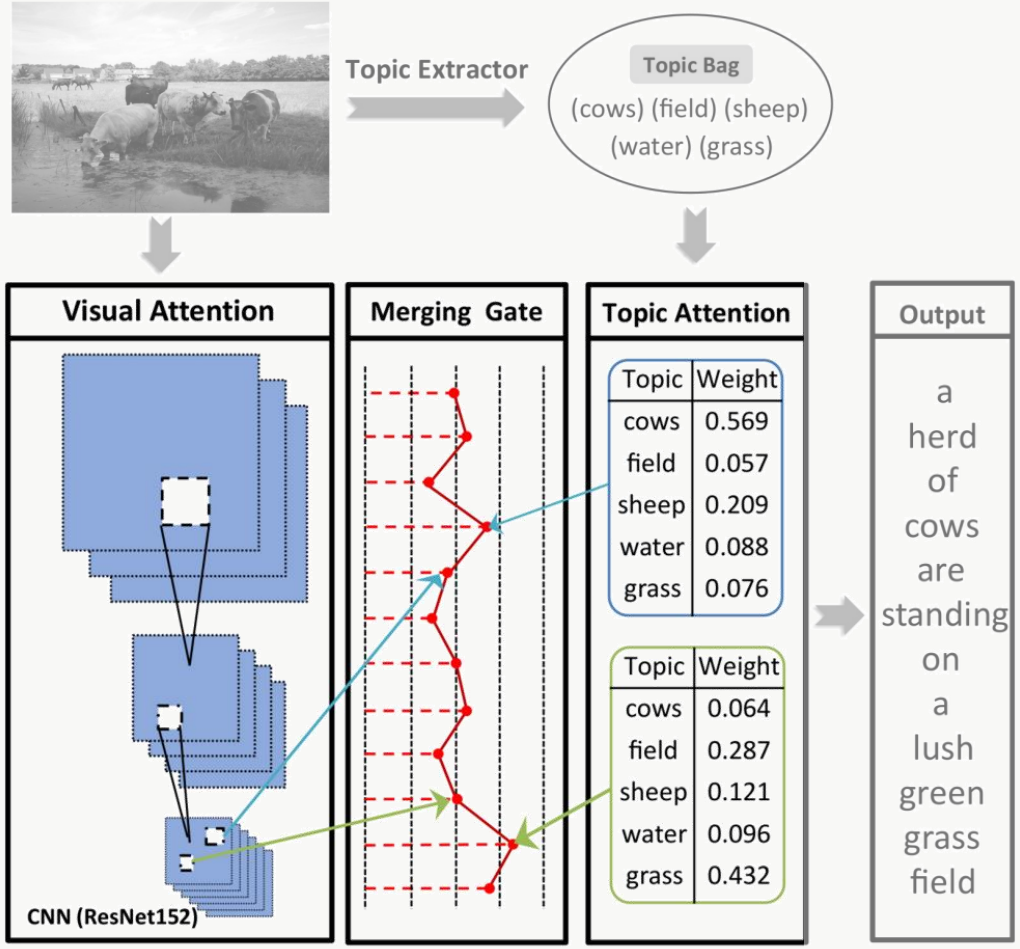
Visual Attention: CNN(ResNet152)을 기반으로 Visual information(r_t) 추출



Topic Attention: Topic Extractor로부터 추출한 후보 Topic으로부터 contextual information(s_t) 생성



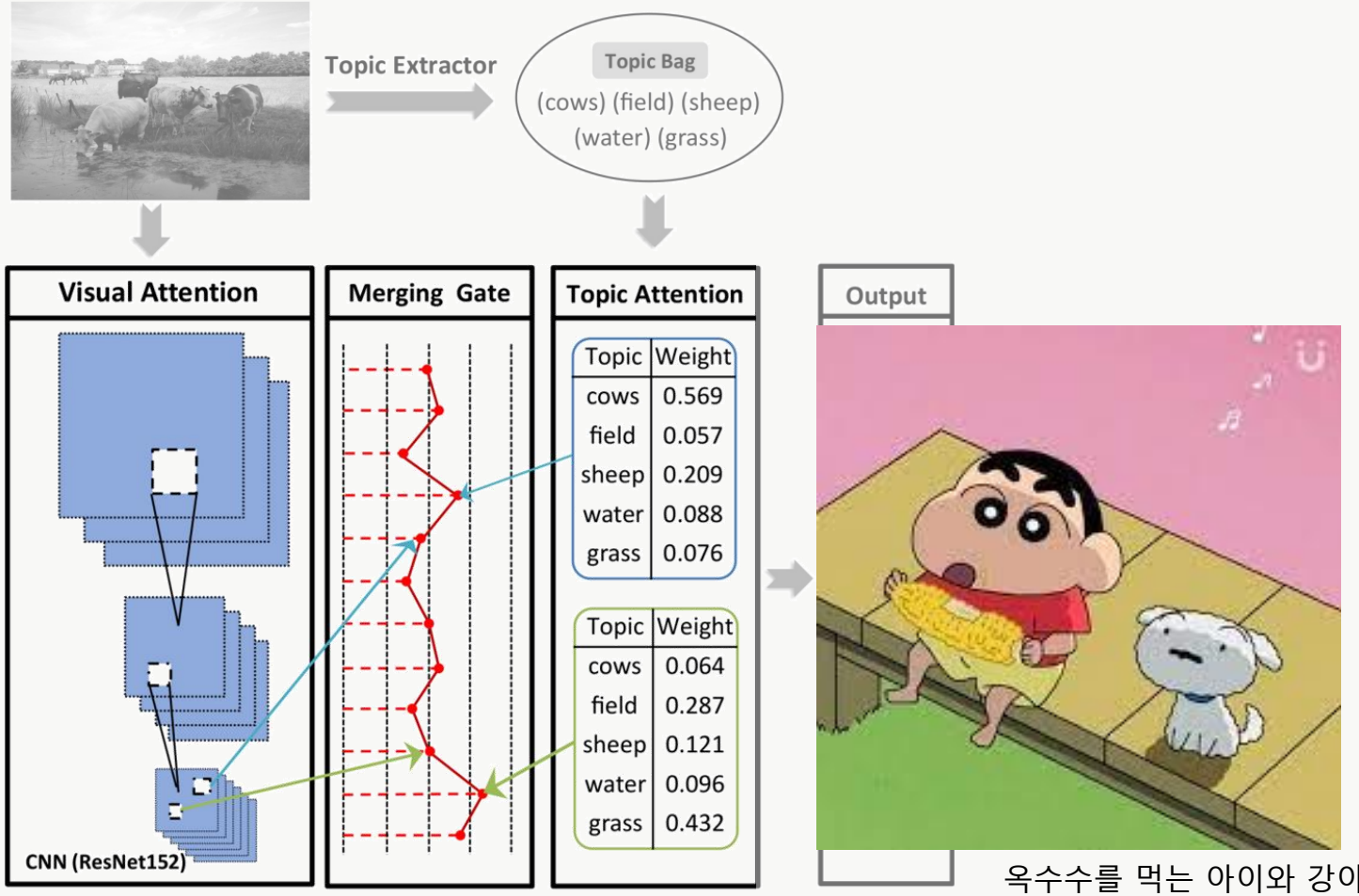
Merging Gate: Visual information(r_t)와 contextual information(s_t) 간 적응적인 조합을 통해 다음에 올 단어에 대한 확률 추출



γ_t: s_t가 r_t보다 얼마나 더 중요한 지

$$\gamma_t s_t + (1 - \gamma_t) r_t$$

Merging Gate: Visual information(r_t)와 contextual information(s_t) 간 적응적인 조합을 통해 다음에 올 단어에 대한 확률 추출

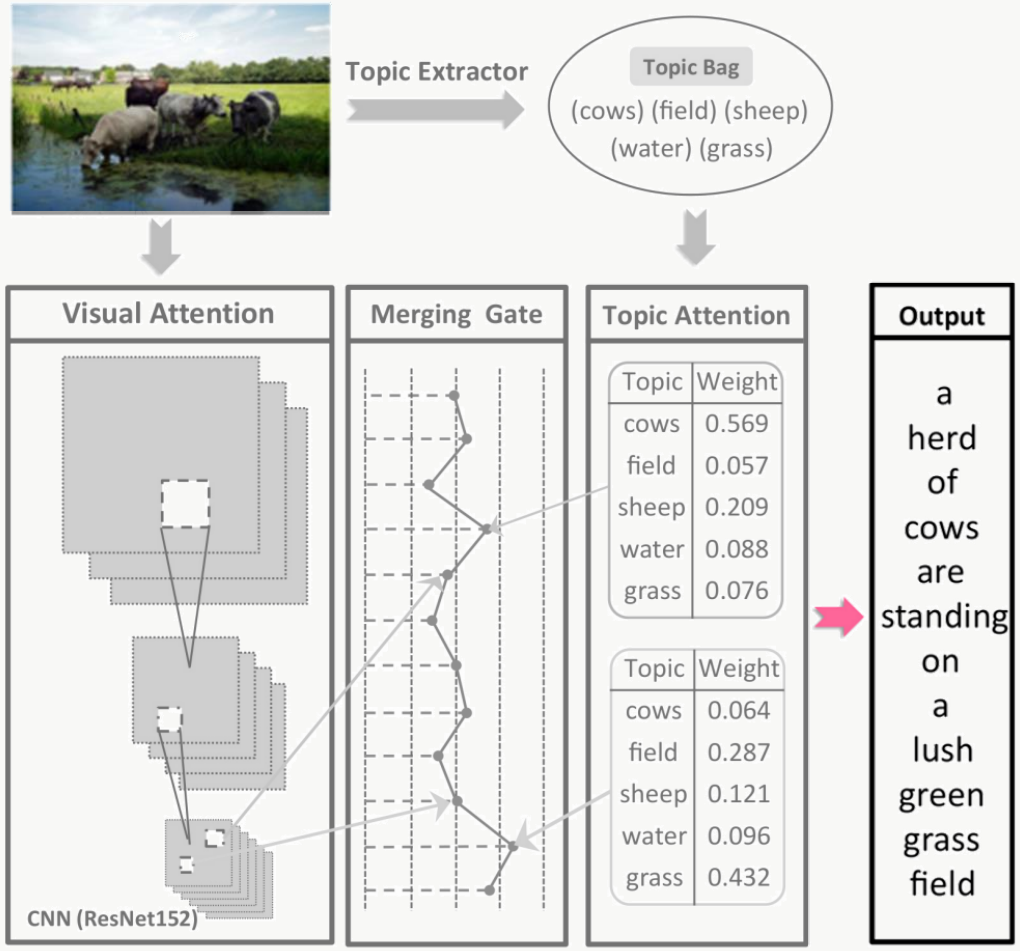


옥수수를 먹는 아이와 강아지
옥수수 vs 먹는

γ_t: s_t가 r_t보다 얼마나 더 중요한 지

$$\gamma_t s_t + (1 - \gamma_t) r_t$$

Merging Gate: Visual information(r_t)와 contextual information(s_t) 간 적응적인 조합을 통해 다음에 올 단어에 대한 확률 추출



확률을 기반으로 단어를 생성하여 최종 Output 출력

Contents

Search

03 Conclusion (1) 평가

Search

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

SOTA 방법에서 평가 점수가 높은 편에 속하며
인간의 평가와 상관관계가 가장 높은 평가 지표인 **SPICE**에서 가장 높은 점수를 달성

03 Conclusion (2) 분석

Search

Comparison of Models					
Topics	woman girl baby bear kitchen	computer keyboard laptop mouse desk	buildings bus clock tower street	pizza cheese table plate toppings	motorcycle street car bike motorcycles
Visual Attention	a girl and a baby are holding a stuffed animal	a computer keyboard sitting on top of a wooden desk	two green buses is parked on the side of the road	two pizzas with toppings on a table	a row of motorcycles parked next to each other
Topic Attention	a woman holding a teddy bear in a kitchen	a computer keyboard and a mouse sitting on a desk	a large double decker bus is parked in front of a building	a pizza with a lot of toppings on it	a motorcycle parked in a parking lot next to a car
simNet	a woman and a baby are holding a stuffed animal	a computer keyboard and mouse on a wooden desk	two green double decker buses parked in front of a large building	two pizzas sitting on a table with two different kinds of toppings	a row of motorcycles parked in a street

Error Analysis			
Topics	clock tower building street city	people bus truck street train	garden bench park forest plants
Reference	a tall building that has a clock on it (near a large building)	tour buses driving down a street lined with cheering people	an old wooden bench in nature surrounded by plants
simNet	a large building with a clock tower in the background	a group of people standing around a parked bus at a bus stop	a wooden bench sitting in the middle of a lush green garden
Error Type	distance	movement	object

03 Conclusion (2) 분석

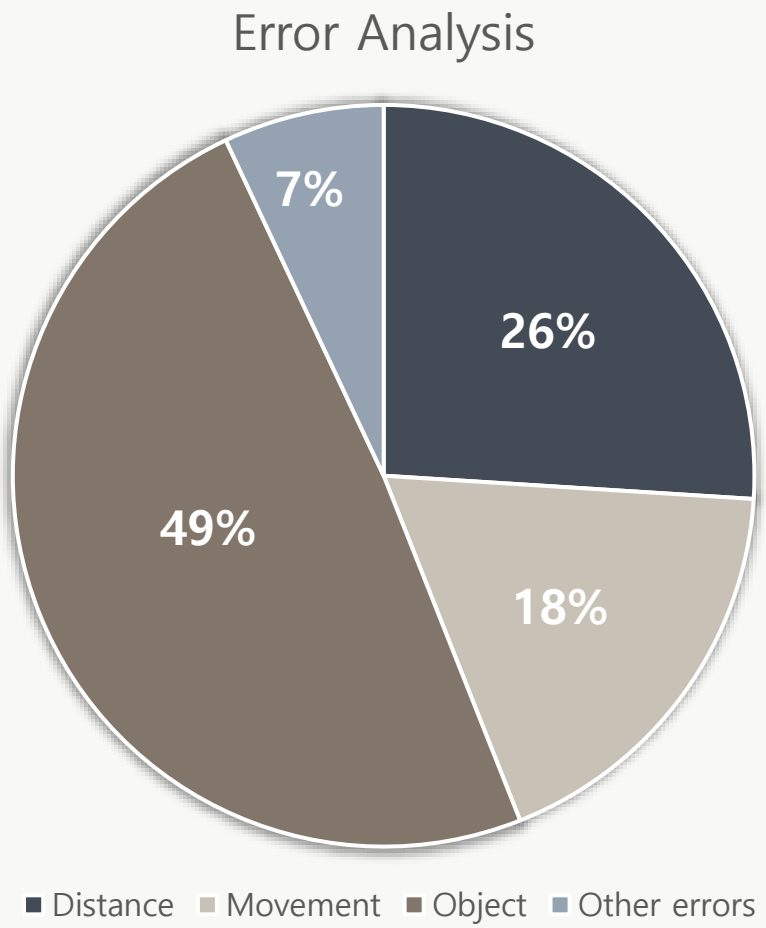
Search

Comparison of Models					
Topics	woman girl baby bear kitchen	computer keyboard laptop mouse desk	buildings bus clock tower street	pizza cheese table plate toppings	motorcycle street car bike motorcycles
Visual Attention	a girl and a baby are holding a stuffed animal	a computer keyboard sitting on top of a wooden desk	two green buses is parked on the side of the road	two pizzas with toppings on a table	a row of motorcycles parked next to each other
Topic Attention	a woman holding a teddy bear in a kitchen	a computer keyboard and a mouse sitting on a desk	a large double decker bus is parked in front of a building	a pizza with a lot of toppings on it	a motorcycle parked in a parking lot next to a car
simNet	a woman and a baby are holding a stuffed animal	a computer keyboard and mouse on a wooden desk	two green double decker buses parked in front of a large building	two pizzas sitting on a table with two different kinds of toppings	a row of motorcycles parked in a street

03 Conclusion (2) 분석

Search

Error Analysis			
Topics	clock tower building street city	people bus truck street train	garden bench park forest plants
Reference	a tall building that has a clock on it (near a large building)	tour buses driving down a street lined with cheering people	an old woo den bench in nature surrounded by plants
simNet	a large building with a clock tower in the background	a group of people standing around a parked bus at a bus stop	a wooden bench sitting in the middle of a lush green garden
Error Type	distance	movement	object



03 Conclusion (3) 결론

Search

1

Visual
+
Conceptual
병합 네트워크
최초 제안

2

stepwise
merging
mechanism
도입

3

SPICE 측면
에서 SOTA
성능 능가

감사합니다

Search