

[Paper Review]

ConZIC: Controllable Zero-shot Image Captioning by Sampling-Based Polishing

CVPR 2023

Index

1. Introduction
2. Related work
3. Controllable Zero-shot Image Captioning, ConZIC
4. Method
5. Experiments
6. Conclusion and future work

Introduction

❖ 최근 제로샷 가능성은 딥러닝에서 중요한 이슈이다.

- Supervised methods 의 한계점

- 많은 양의 high quality paired data 에 의존한다.
- train data 분포에서 벗어나는 real-world 정보 반영 어려움

→ Zero-shot image captioning, **ZeroCap**

: 거대 pre-trained 모델의 지식을 활용하여 지도학습 없이 캡션을 생성한다.

- 그러나, ZeroCap에도 한계점이 존재한다.

- Zerocap의 autoregressive generation 방법은 캡션의 다양성을 제한
- Zerocap의 gradient-directed searching 방법은 추론 속도를 제한
- 또한, zero-shot image captioning에서의 controllability에 대한 고려는 반영되지 않음

Introduction

본 논문에서는,

- 1) Gibbs sampling과 MLM의 관계를 분석하여, 새로운 Language Model인 **Gibbs-BERT**를 제안하고,
- 2) 이를 CLIP과 결합한 **Controllable Zero-shot Image Captioning method**인 **ConZIC**을 소개한다.

By using Gibbs-BERT,

언어모델이 Sampling-based search를 통해 더 자유로운 생성 순서를 갖는다.

- 더 빠르고 다양한 캡션 생성 가능해짐

By integrating with CLIP,

이미지와 텍스트 사이의 유사도를 반영한다.

- Zero-shot Image Captioning을 수행한다.

By adding Task-specific discriminator,

Task에 맞는 Controllable Image Captioning을 수행한다.

Related work

❖ Supervised Image Captioning

- 이미지 캡셔닝의 많은 previous work는 Supervised 방법으로 연구되어 왔다.

- CNN-based encoder와 RNN-based decoder 결합한 초기 모델
 - ex) Show and tell
- Attentive Object detector 사용 모델
 - ex) BUTD, M2 ..
- Graph neural network 사용 모델
 - ex) Scene Graph Auto Encoding
- 거대 Visual-Language Pretrain 모델
 - 최근 많은 downstream task에서 SOTA 달성
 - ex) VinVL, Oscar, LEMON..

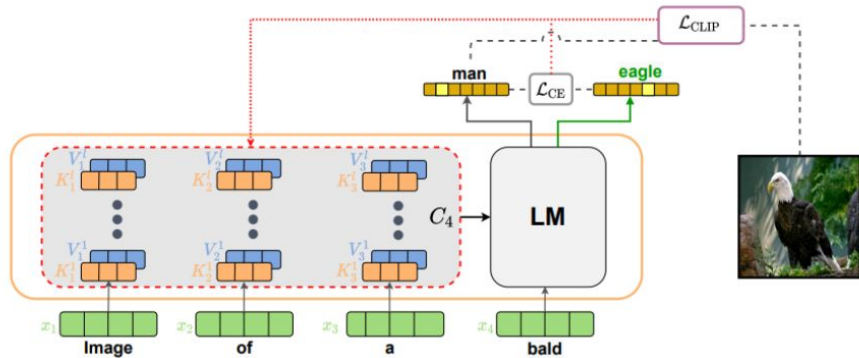
→ 큰 성능 발전이 있었지만,
여전히 supervised fine-tuning 작업이 필요하다.

Related work

❖ Zero-shot Image Captioning

- 최근 거대 pre-trained model 활용한 zero-shot 연구가 진행되어, training data 없이도 task 수행이 가능하게 되었다.

- ZeroCap : 거대 pt 모델의 지식을 활용하여 지도학습 없이 캡션을 생성
 - CLIP으로부터 학습된 V-L knowledge를 활용하여 계산된 clip score로 image-text 매칭을 유도하고,
 - GPT-2로부터 학습된 linguistic knowledge를 활용하여 자연스러운 캡션 생성하게끔 한다.



ZeroCap은 Optimization 과정에서 context cache $C_i = [(K_j^l, V_j^l)]_{j < i, 1 \leq l \leq L}$ 를 매 time step 마다 조정한다.

⇒ 즉, Optimization은 Autoregression 과정 동안 이루어지며, 각 토큰마다 반복된다. (반복적인 gradient updating)

Related work

❖ Diversity Aspect

1) Supervised methods

- "mode collapse" 문제 발생
: 생성된 결과가 평균에 편향되어 있음을 의미한다.
이로 인해 다양한 단어와 문장패턴 (구문론적 / 의미론적 다양성) 을 구성하지 못한다.

2) Zero-shot method (ZeroCap)

- Vocab의 다양성은 증가했으나, **(more semantical diversity) (+)**
- Autoregressive decoding 방식은 의미론적으로 비슷한 패턴을 가진 문장을 생성한다. **(less syntactical diversity) (-)**

=> 구문론적 / 의미론적 diversity 모두 고려한 image captioning 연구가 필요하다.

❖ Controllability Aspect

- Control signal을 image captioning에 적용한 이전 연구도 존재한다.
 - Subjective signals : sentiments, emoticons, personality, ..
 - Objective signals : length level, parts-of-speech, object region, visual relation, ..

=> Control signal을 **zero-shot** image captioning에 적용한 연구가 필요하다.

Controllable Zero-shot Image Captioning, ConZIC

- ❖ 앞서 소개한 한계점을 극복하는 **Controllable Zero-shot Image Captioning, ConZIC**을 소개한다.

1. More Flexible

- Autoregressive generation은 한방향으로 생성되므로, 단어가 한 번 생성 되면 바꿀 수 없다. -> not flexible
- Gibbs-BERT는 **bidirectional**하게 정보를 반영하고, 초기 생성된 캡션을 더 나은 방향으로 수정할 수 있다. -> more flexible

2. More Efficient

- ZeroCap의 반복적인 context cache update는 계산 시간을 증가시킨다. -> inefficient
- ConZIC은 추가 parameter update 과정 없기 때문에, **ZeroCap보다 5배 빠른 속도** -> efficient

Controllable Zero-shot Image Captioning, ConZIC

3. More Diverse

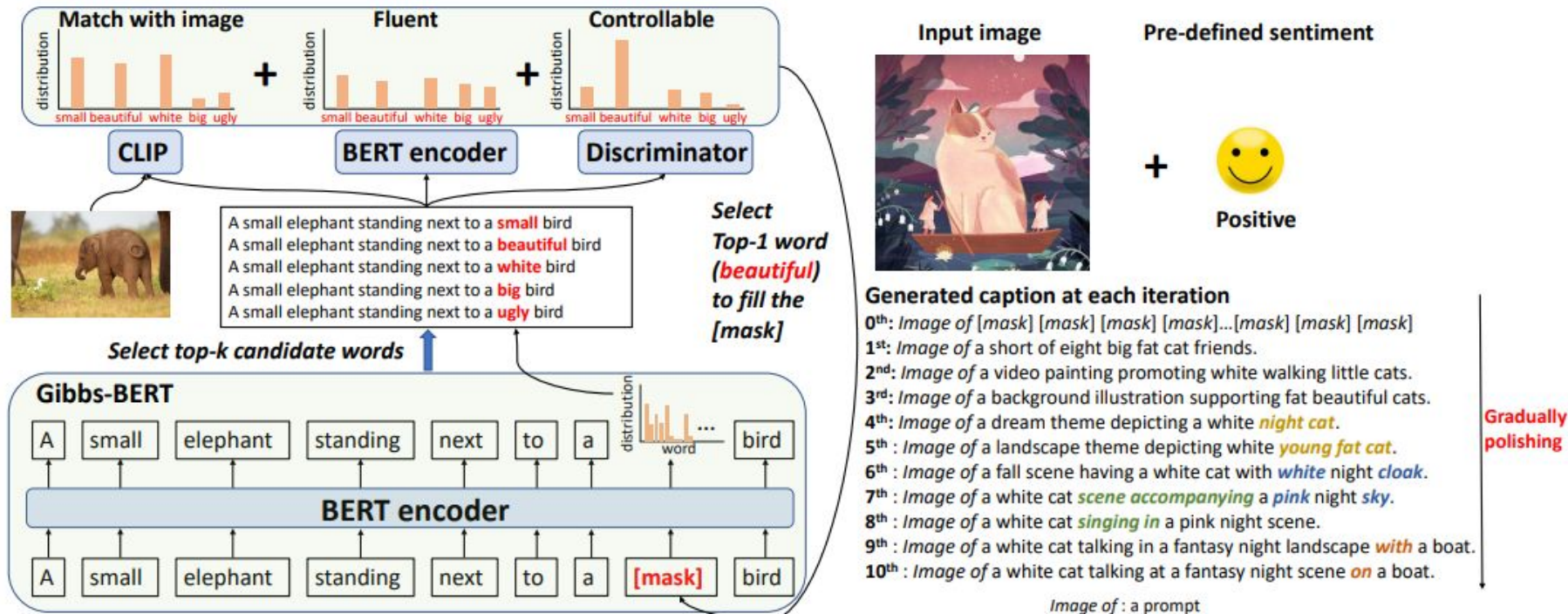
- ZeroCap은 beam search로 후보 문장을 생성하므로 (deterministic method), 캡션이 유사한 패턴을 보인다. -> **low diversity**
- ConZIC의 Gibbs-BERT는 generation 과정에서 flexible하게 searching 수행한다. -> **higher diversity**

4. More Controllable

- 네 개의 **controllable signals** (length, infilling, styles, and parts-of-speech)을 사용한다.
- 최초의 controllable zero-shot Image Captioning method 이다.

Controllable Zero-shot Image Captioning, ConZIC

ConZIC의 전반적인 아키텍처는 다음과 같다.



Controllable Zero-shot Image Captioning, ConZIC

❖ Framework of ConZIC

ConZIC은 이미지 I 와 Control signal C 가 주어졌을 때,

언어 모델의 Likelihood $p(\mathbf{x}_{<1,n>}|I, C)$ 를 최대화 하는 $\mathbf{x}_{<1,n>}$ 를 searching하며 캡셔닝을 수행하고,

목적함수를 수식으로 나타내면 다음과 같다.

$$\begin{aligned} & \log p(\mathbf{x}_{<1,n>}|I, C) \\ \propto & \log p(\mathbf{x}_{<1,n>}, I, C) \\ = & \log p(I|\mathbf{x}_{<1,n>}) + \log p(C|\mathbf{x}_{<1,n>}) + \log p(\mathbf{x}_{<1,n>}), \end{aligned} \tag{1}$$

$\mathbf{x}_{<1,n>}$: n 개의 단어로 이루어진 문장

I : Image

C : Control signal

- 1) $\log p(\mathbf{x}_{<1,n>})$: 자연스러운 문장 생성을 유도한다
- 2) $\log p(I|\mathbf{x}_{<1,n>})$: image 와 캡션의 유사도를 계산한다
- 3) $\log p(C|\mathbf{x}_{<1,n>})$: control signal을 만족하는 캡션 생성하도록 유도

→ 반복적으로 샘플링을 수행하며 더 나은 캡션을 생성한다. (Polishing)

Method

❖ Method

- 기존 Autoregressive Generation $p(\mathbf{x}_{<1,n>}) = p(x_n|\mathbf{x}_{<n>}) \cdots p(x_2|x_1)p(x_1)$

이전까지의 토큰을 고려해 다음 토큰을 autoregressive하게 생성한다.

한계 : low diversity, error accumulation

=> 본 논문에서는 **Gibbs sampling** 과 **Masked Language model** 을 결합한 'sampling-based LM' 을 새롭게 정의한다.

* Gibbs sampling ?

- MCMC 알고리즘에 기반한 샘플링 방법
: data의 결합분포 $p(\mathbf{x}_{<1,n>})$ 로부터 반복적으로 x_i 샘플링하고, 반복을 통해 조건부분포 $p(x_i|\mathbf{x}_{-i})$ 에 가까운 x 를 샘플링한다.
(\mathbf{x}_{-i} : x_i 제외한 다른 랜덤 변수)
- 샘플링 초기에는 \mathbf{x}_{-i} 에 크게 의존하지만, 충분히 많이 뽑고 난 뒤에는 초기 상태에 관계없이 p 에 기반한 표본 수집할 수 있다.
- 또한, Gibbs sampling은 샘플링 순서가 자유롭다.

→ mode collapse 를 벗어나 더 다양한 캡션 생성을 가능하게 한다.

Method

• ConZIC의 Masked Language Model

- Gibbs sampling 을 MLM에 적용
- x_M 이 [MASK] 토큰 일때, MLM은 다른 단어들 x_{-M} 로부터 x_M 이 나올 확률 분포를 학습하는 것이 목표이고, 이는 Gibbs sampling에서 $p(x_i|x_{-i})$ 를 예측하는 과정과 동일하게 볼 수 있다.

Algorithm 2: Algorithm of Gibbs-BERT.

Data: initial sentence: $\mathbf{x}_{<1,n>}^0 = (x_1^0, \dots, x_n^0)$;

iterations= T , candidates= K ;

position sequence $P = \text{Shuffle}([1, \dots, n])$;

Result: the final sentence: $\mathbf{x}_{<1,n>}^T = (x_1^T, \dots, x_n^T)$;

for iteration $t \in [1, \dots, T]$ **do**

 state: $\mathbf{x}_{<1,n>}^{t-1} = (x_1^{t-1}, \dots, x_n^{t-1})$;

for position $i \in P$ **do**

 1. Replace x_i^{t-1} with [MASK];

 2. Predict the word distribution over vocabulary
 by BERT: $p(x_i|\mathbf{x}_{-i}^{t-1})$;

 3. Sample x_i from distribution $p(x_i|\mathbf{x}_{-i}^{t-1})$;

 4. Replace x_i^{t-1} with x_i^t ;

end

 state: $\mathbf{x}_{<1,n>}^t = (x_1^t, \dots, x_n^t)$;

end

x_i^t 예측하는 과정

- x_i^{t-1} = [MASK] 로 마스킹
- 매 반복마다, BERT의 word distribution으로 부터 word 분포를 예측하고, 그 분포를 따르는 단어를 샘플링
- [MASK] 대신 샘플링 된 값 대입
- $x_i^t = x_i^{t-1}$ 로 대체
- 현재 step에서 구한 token은 다음 iteration에 반영됨
- 반복을 통해 Polishing

Method

2. Image-text matching network for p_k^{Clip}

$$p(I|\{s_k\}_{k=1}^K) \propto \text{Softmax}[CLIP(s_k, I)].$$

- image-text 유사도 반영하기 위해 **CLIP matching score** $CLIP(s_k, I)$ 를 계산한다.

- Gibbs-BERT가

1) top-K 후보 단어 선정하고,

2) i 번째 [MASK] token 대신 K개의 후보 단어를 넣고,

K개의 후보 문장을 생성한다. $\{s_k = (x_1, \dots, x_{ik}, \dots, x_n)\}_{k=1}^K; x_{ik} = [MASK]$

3) 후보 문장들과 이미지 간의 CLIP 유사도를 구하고, softmax 취해 최종 **CLIP matching score**를 얻는다.

Method

3. Discriminator for control signal, p_k^{Cls}

$$p(C|\{s_k\}_{k=1}^K) \propto \text{Softmax}[\text{Classifier}(s_k)]$$

- task 에 맞는 pre-trained classifier를 사용하여 후보 문장에 대한 controllable score 를 구한다.
 - $p(C|x_{<1,n>})$ 없으면 일반적인 zero-shot Image Captioning을 수행한다.
 - length control 같은 특정 controllable task에서는 $p(C|x_{<1,n>})$ 사용할 필요 없다.
(style, pos와 같은 task에는 Pre-trained Classifier 필요)

→ 앞서 언급한 framework $\log p(I|x_{<1,n>}) + \log p(C|x_{<1,n>}) + \log p(x_{<1,n>})$ 에 따라,

최종적인 단어 예측 확률은 $\alpha p_k^{Bert} + \beta p_k^{Clip} + \gamma p_k^{Cls}$ 이다.

Method

Algorithm 1: Algorithm of our proposed ConZIC.

Data: initial caption: $\mathbf{x}_{<1,n>}^0 = (x_1^0, \dots, x_n^0)$;

iterations= T , candidates= K ;

position sequence $P = \text{Shuffle}([1, \dots, n])$;

Result: the final caption: $\mathbf{x}_{<1,n>}^T = (x_1^T, \dots, x_n^T)$;

for iteration $t \in [1, \dots, T]$ **do**

 state: $\mathbf{x}_{<1,n>}^{t-1} = (x_1^{t-1}, \dots, x_n^{t-1})$;

for position $i \in P$ **do**

 1. Replace x_i^{t-1} with $[MASK]$;

 2. Predict the word distribution over vocabulary

 by Gibbs-BERT: $p(x_i | \mathbf{x}_{-i}^{t-1})$;

 3. Select top- K candidate words $\{x_{ik}^t\}_{k=1}^K$ by
 $p(x_i | \mathbf{x}_{-i}^{t-1})$, whose probability is p_k^{Bert} ;

 4. Get K candidate sentences $\{s_k\}_{k=1}^K$:
 $(x_1^{t-1}, \dots, x_{i-1}^{t-1}, x_{ik}^t, x_{i+1}^{t-1}, \dots, x_n^{t-1})_{k=1}^K$;

 5. Compute the CLIP and classifier score for
 $\{s_k\}_{k=1}^K$ by Eq. 4 and 5: p_k^{Clip} and p_k^{Cls} .

 6. Select x_i^t with largest probability by
 $\alpha p_k^{Bert} + \beta p_k^{Clip} + \gamma p_k^{Cls}$;

 7. Replace x_i^{t-1} with x_i^t ;

end

 state: $\mathbf{x}_{<1,n>}^t = (x_1^t, \dots, x_n^t)$;

end

• Overall Algorithm

Process

-> 초기 캡션 : "Image of [MASK] [MASK] [MASK]..."

-> Gibbs-BERT가 top-K 후보 단어 선정하여 K개의 후보 문장 s_k 선별

-> 후보 문장에 대한

text-image score $p(\text{I} | \mathbf{x}_{<1,n>})$ & text-control matching score $p(\text{C} | \mathbf{x}_{<1,n>})$ 얻음

-> 이를 Gibbs-BERT predicted distributions $p(x_i | \mathbf{x}_{-i})$ 와 결합해 최종 확률분포 구함

$$= \text{Final distribution} \quad \alpha p_k^{Bert} + \beta p_k^{Clip} + \gamma p_k^{Cls}$$

-> 확률값 가장 높은 x_i 선택

Experiments

Settings

- Datasets : MSCOCO caption, SentiCap, FlickrStyle10k, SketchyCOCO caption
- Image-Text matching network : **CLIP-ViT-B/32**
- Language Model : **BERT-Base**
=> 추가 fine tuning 진행하지 않음
- $K, T, \alpha, \beta, \gamma = 200, 15, 0.02, 2, 5$ (iterations = T , candidates = K)

Evaluation Metrics

- Evaluate accuracy
: BLEU-4 (B-4), METEOR (M), CIDEr (C), SPICE (S), RefCLIPScore (RefCLIP-S)
- Unsupervised metric (semantic-related)
: CLIPScore(CLIPS) -> 이미지와 생성된 캡션간 유사도 측정
- Evaluate diversity
: Vocab, Self-CIDEr(S-C), Div-n -> 단어의 다양성, 캡션들끼리의 유사도 측정

Experiments

1. Standard Image Captioning task

	Accuracy						Diversity			
Metrics			Supervised			Unsupervised				
	B-4(↑)	M(↑)	C(↑)	S(↑)	RefCLIP-S(↑)	CLIP-S(↑)	Vocab (↑)	S-C(↑)	Div-1(↑)	Div-2(↑)
Supervised Methods										
ClipCap [49]	32.15	27.1	108.35	20.12	0.81	0.77	1650	-	-	-
MAGIC [64]	12.90	17.22	48.33	10.92	0.77	0.74	1765	-	-	-
CLIP-VL [61]	40.2	29.7	134.2	23.8	0.82	0.77	2464	-	-	-
ViTCAP [22]	41.2	30.1	138.1	24.1	0.80	0.73	1173	-	-	-
GRIT [50]	42.4	30.6	144.2	24.3	0.82	0.77	1049	-	-	-
VinVL [80]	41.0	31.1	140.9	25.2	0.83	0.78	1125	-	-	-
LEMON [33]	42.6	31.4	145.5	25.5	-	-	-	-	-	-
Supervised and Diversity-based Methods										
Div-BS [72]	32.5	25.5	103.4	18.7	-	-	-	-	0.20	0.25
AG-CVAE [68]	31.1	24.5	100.1	17.9	-	-	-	-	0.23	0.32
POS [19]	31.6	25.5	104.5	18.8	-	-	-	-	0.24	0.35
ASG2Caption [14]	31.6	25.5	104.5	18.8	-	-	-	0.76	0.43	0.56
Zero Shot Methods										
ZeroCap [65]	2.60	11.50	14.60	5.50	0.79	0.87	8681	0.63	0.31	0.45
Ours (sequential)	1.31	11.54	12.84	5.17	0.83	1.01	9566	0.63	0.40	0.56
Ours (shuffle)	1.29	11.23	13.26	5.01	0.83	0.99	15462	0.95	0.62	0.87

Experiments

1. Standard Image Captioning task

Results

- BLEU-4, METEOR, CIDEr, SPICE metrics에서 supervised method 보다 성능 좋지 못함
이유 : supervised metrics이므로 MSCOCO의 train, test data annotation에서의 유사성이 반영됨
train과 test에서 비슷한 캡션 스타일로 인해 **domain bias**가 생긴다.
=> 생성된 캡션의 다양성 떨어짐
- Diversity 측면에서 기존 Diversity-based Methods 보다 큰 차이로 좋은 성능을 보인다.
- Semantic-related metrics 에서도 좋은 성능을 보인다.

Experiments

- Zero-shot performance

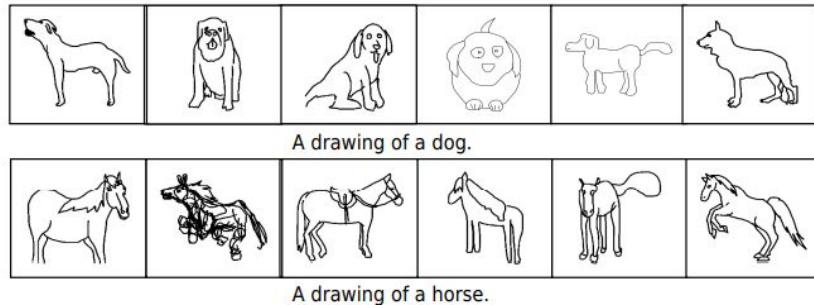
zero-shot performance를 측정하기 위해,

MSCOCO로 train된 supervised method를 SketchyCOCO caption dataset으로 성능 평가를 진행했다.

Methods		B-1(↑)	M(↑)	C(↑)	CLIP-S(↑)
Supervised	MAGIC [64]	21.88	11.77	13.00	0.66
	ViTCAP [22]	27.69	17.58	22.29	0.63
	GRIT [50]	17.84	26.62	17.84	0.68
Zero Shot	ZeroCap [65]	27.08	20.67	21.11	0.86
	Ours	39.61	20.71	34.43	0.88

Table 2. Performance on SketchyCOCO caption dataset.

SketchyCOCO caption dataset 예시



- ConZIC이 Supervised method 보다 나은 성능을 보인다.
이유: Supervised method는 MSCOCO 와 SketchyCOCO 간 domain gap 존재 -> **generality 떨어짐**
- ConZIC은 ZeroCap 보다도 나은 성능을 보였다.

Experiments

- Qualitative Results



GRIT: A painting of a painting with a tree in the background

CLIPCap: The night sky over the city.

ViTCap: A painting of a bird on a table with a bird on it.

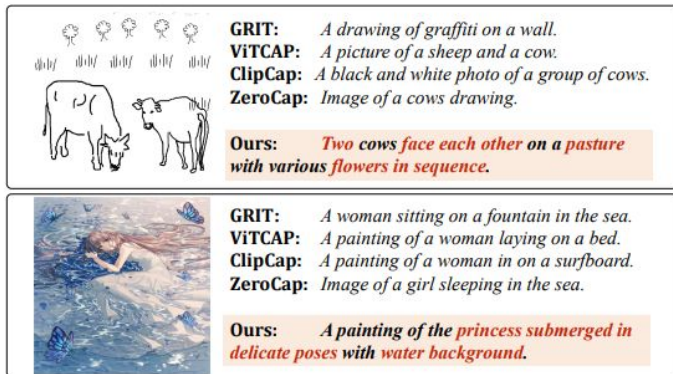
ZeroCap: A night with Vincent.

Ours:

A famous Gogh painting after streaming moonlight over all the grand structures.

A view despite a nocturnal sky within famous mainstream artworks.

A nighttime sky can appear in drawings and oil paintings.



(a) Examples of zero-shot image captioning.

- 다양한 스타일의 이미지에서도 정확하고 다양한 캡션을 생성한다.
- Shuffling the word generation order 로 인해 Zerocap의 beam search 보다 구문론적, 의미론적으로 다양한 결과 생성한다.


Experiments

2. Controllable Image Captioning tasks

다음으로, 네 가지 Controllable tasks에 대한 실험을 진행했다.

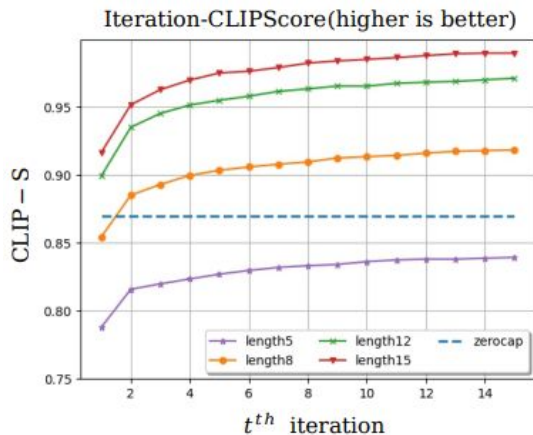
❖ Length

- 4개의 서로 다른 length로 실험 (5, 8, 12, 15)
- classifier 추가 없이 초기 length값만 지정해준다.

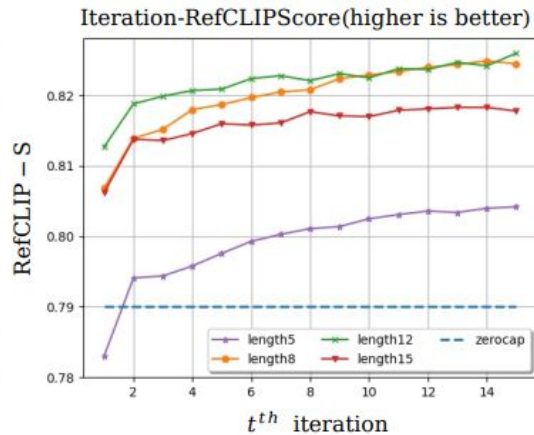
sentence length control					
	3~5	A stuffed black bear.	A fruit dish.	A calm businessman concentrating hard.	A blond farm cow.
	7~9	A bear toy named Cooper admiring himself.	A fruit dish in tin color offering sweet orange.	A financial administrator watching financial statements online.	A farm buffalo around metal enclosure and foliage.
	11~13	A stuffed teddy dark bear smiling with yoga pose in a mirror.	A photo showing Osaka orange fruits appearing in a stainless steel pot.	A man distracted thinking business report with neatly trimmed white hair.	A village animal cow shows in tree ferns background and fences.
					A cornered cat shown against numerous pigeons.
					A mute cat meeting numerous birds and pigeons in a Greek square.

Experiments

❖ Length



(a) Comparison on CLIPScore.



(b) Comparison on RefCLIPScore.




- 캡션 길이 길수록 많은 정보 담고 있어 더 높은 CLIPScore, RefCLIPScore 를 보인다.
- length 12, 8 이 5, 15일 때 보다 나은 RefCLIPScore 를 보인다.

이유 : MSCOCO 캡션 평균 length : 10

-> 생성된 캡션의 length가 10에 가까울 때 reference와의 유사도 더 높게 측정

Experiments

❖ Infilling

Infilling			
Reference	Two zebras fighting in a cloud of dust.	A young man wearing black attire and a flowered tie is standing and smiling.	Multiple wooden spoons are shown on a table top.
Corrupted	_____ zebras _____ of dust.	_____ flowered tie _____ black _____ a _____ smiling.	_____ wooden spoons _____ on _____ table top.
Infilling	Male zebras fought amidst the cover of dust.	Kyle was wearing a black suit with a flowered tie perfectly and was smiling.	Many wooden spoons were arranged on the table top.
Corrupted	Two of dust. _____ in a _____	A _____ and _____ tie is standing and smiling.	_____ are _____ on a table top.
Infilling	Two female zebras compete in a patch of dust.	A goth teen with cracked glasses and a thick Asian tie is standing and smiling.	Wooden art spoons are organized on a table top.
Corrupted	Two _____ a _____ of dust.	A young _____ attire and a _____ standing _____ smiling.	Multiple wooden _____ are _____ on _____ table top.
Infilling	Two zebras having provoked a disk of dust.	A young teen in business attire and a black glasses is standing and smiling.	Multiple wooden breakfast forks are stored on the table top.

- 빈칸을 채우기 위해 image 정보와 고정된 주변 단어 정보를 모두 반영한 결과이다.
- ConZIC의 Gibbs-BERT => bidirectional attention에 기반한 모델이므로 classifier를 추가하지 않고도 task 수행가능 (Zerocap은 autoregressive LM 사용하므로 해당 task 수행시 left context 정보만 반영)

Experiments

❖ Infilling

• One-word-infilling (한 단어 빈 칸)

* WSim, BSim ?

: 생성 단어와 reference 단어 간 유사도 측정하는 방법

WSim : measures node distance in Wordnet

BSim : cosine distance in BERT word embedding space

결과 : ConZIC이 Zerocap보다 나은 성능을 보인다.

• Multiple-word-infilling (여러 단어 빈 칸)

결과 : 빈칸의 비율 높을수록 CLIP-S은 높아지고, B-4, M은 낮아진다.

Method	Noun			Verb		
	B-1(↑)	WSim(↑)	BSim(↑)	B-1(↑)	WSim(↑)	BSim(↑)
ZeroCap [65]	0.00	0.001	0.11	0.00	0.001	0.10
ConZIC	0.37	0.39	0.52	0.25	0.46	0.50

Table 3. Results of one word infilling task on MSCOCO dataset.

Corrupted Ratio	B-4(↑)	M(↑)	CLIP-S(↑)
0.25	60.69	44.99	0.83
0.50	26.08	29.12	0.89
0.75	8.06	17.60	0.93

Table 8. Results of multiple-words-infilling task.

Experiments

❖ Style

- SentiCap
: positive, negative, romantic, or humorous 같은 스타일 반영한 데이터셋
- Pre-trained classifier가 필요하다.
 - SentiwordNet : sentiment (positive or negative) controlling
 - TextCNN : romantic-humorous controlling




		Positive			
Metrics		B-3(↑)	M(↑)	CLIP-S(↑)	Acc(↑)
Supervised	StyleNet [24]	12.1	12.1	-	45.2
	MSCap [30]	16.2	16.8	-	92.5
	MemCap [81]	17.0	16.6	-	96.1
Zero Shot	ConZIC	1.89	5.39	0.99	97.2
		Negative			
Supervised	StyleNet [24]	10.6	10.9	-	56.6
	MSCap [30]	15.4	16.2	-	93.4
	MemCap [81]	18.1	15.7	-	98.9
Zero Shot	ConZIC	1.78	5.54	0.97	99.1

• 결과

- B-3 and M 에서 기존 supervised SOTA method 성능이 더 좋았지만,
- 생성 캡션을 sentiment classify 하여 구한 acc 결과는 ConZIC이 더 높음

Experiments

❖ Parts-of-speech (POS)

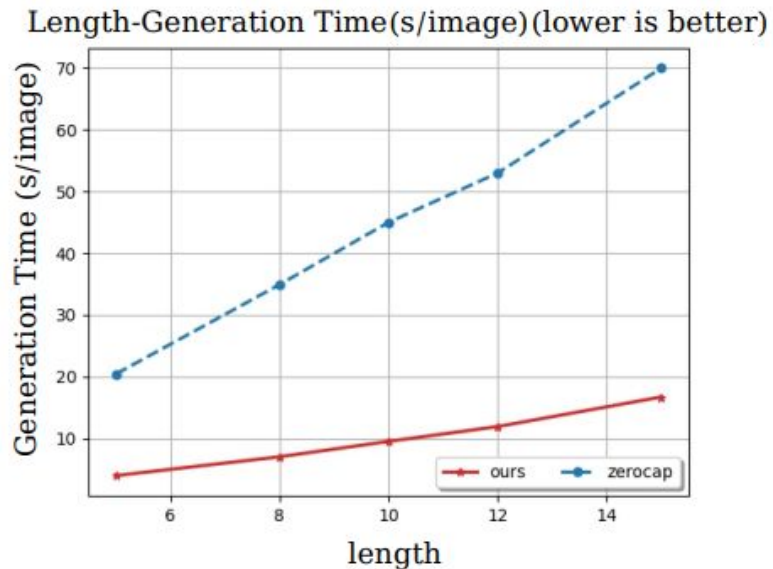
Parts-of-speech (POS) control			
POS	DET ADJ/NOUN NOUN VERB VERB ADV ADP DET ADJ/NOUN NOUN NOUN.		
Without control	A gray dog embracing a familiar sad purple elephant with shark clothing.	A darkened London lane with a red and gray southbound bus pair.	A female group members in a Portuguese bar counter tasting sparkling wines.
POS control	The grey dog is embraced unexpectedly by a harmless purple elephant.	The electric buses can be seen on a busy Sydney night.	Some female guiders are encountered together during Brazilian wine bar classes.

Parts-of-speech	M(↑)	C(↑)	CLIP-S(↑)	Acc(↑)
without POS	11.54	12.84	1.01	15.54
with POS	7.99	9.29	0.95	86.20

- POS tag 템플릿 [DET ADJ/NOUN NOUN VERB VERB ADV ADP DET ADJ/NOUN NOUN NOUN] 적용 결과, 높은 accuracy 달성했으나, METEOR (M), CIDEr (C), CLIPScore (CLIP-S) 점수는 낮아진 것을 확인할 수 있다. 이유 : 모든 이미지가 POS tags 템플릿에 맞지는 않는다.

Experiments

- Evaluation on Generation Speed



- Sentence length가 speed에 영향을 주는 것을 확인할 수 있다.
- Zerocap에 비해 5배 빠른 추론 속도 (15 iterations 기준)

Conclusion and future work

❖ Contribution

- ConZIC은 **Masked Language Model**과 **Gibbs sampling**의 관련성을 파악하여 새로운 sampling-based language model인 **Gibbs-BERT**를 정의했다.
 - Bidirectional attention을 통해 생성 순서가 자유롭고, 다양한 캡션 생성이 가능해짐
 - 반복적인 sampling을 통해 더 나은 문장을 생성한다.
- **CLIP 기반 image text matching과, pre-trained discriminator**를 활용
 - Control signal이 있을 때와 없을 때 모두 인상적인 zero-shot Image Captioning 성능을 보인다.

❖ Future work

- Zerocap, ConZIC과 같은 zero-shot methods는 작은 물체에 대한 캡셔닝 한계가 존재한다.
(오른쪽 그림의 가위를 인식하지 못함)
-> zero-shot captioning에 더 많은 연구가 필요해 보임
- Controllable image captioning에 더 적절한 metrics 적용할 필요가 보인다.

