

[Paper Review]

**Language Models Can See:
Plugging Visual Controls in Text Generation**

arXiv 2022

Index

1. Introduction
2. Background
3. MAGIC (iMAge-Guided text generatIon with CLIP)
4. Experiments
 - Zero-Shot Image Captioning
 - Story Generation
5. Conclusion

Introduction

❖ Pre-train Language Model

- 최근 거대 Pre-train 모델은 NLP task에서 좋은 성능을 달성했다.

ex) GPT-2

- 디코딩을 통해 next token 예측할 시, textual prompt 사용하여 원하는 task에 맞는 output을 출력하도록 한다.

=> LM의 decoding process에 **이미지**가 관여하는 방법에 대해서는 연구되지 않았다.

❖ Image-Text Embedding Model

ex) CLIP, ALIGN

- 대규모 Noisy Image-Text 쌍을 활용하여 contrastive embedding learning을 수행한다.

=> 인상적인 zero-shot 성능 달성했지만, 이미지 기반 텍스트 생성 task로의 적용은 다른 task에 비해 많이 연구되지 않았다.

→ 본 논문에서는 **이미지 정보를 활용하여 decoding을 가이드** 하는 Image Captioning 모델 **MAGIC** 을 소개한다.

Introduction

❖ MAGIC (iMAge-Guided text generation with CLIP)

: CLIP기반의 "magic score"를 추가하여 새로운 decoding scheme을 적용한 Image-Text 모델

Main Contributions

1. 간단한 “plug-and-play” 원리로 zero-shot 이미지 캡셔닝을 수행하여 여러 benchmarks에서 SOTA를 달성
2. Story generation task에서도 뛰어난 성능을 보였다.
3. 불필요한 gradient update 과정이 없어 기존 zero-shot method인 ZeroCap보다 27배 빠른 속도를 보였다.

Background

❖ Supervised approach in Image Captioning

- **CNN - RNN based methods**
: CNN based encoder로 visual features를 추출하고, RNN/LSTM based decoder로 sentence를 출력하는 방법
 - **Attention Mechanism 활용 methods**
: visual - text의 관계를 잘 표현하여 더 풍부하고 적절한 캡션 생성 가능해짐 (ex. BUTD)
 - **Controllable image captioning**
: 추가 annotation을 필요로 하는 control signal을 사용하여 task에 맞는 설명 캡션을 생성 (ex. Senticap)
 - **V-L pre-training methods**
: 거대 dataset으로 pre-train된 모델을 사용하여 visual-textual 표현을 더 잘 학습할 수 있게 됨 (ex. Oscar, ClipCap, LEMON)
- 여전히 image-text paired data에 의존적이라는 한계가 존재한다.

Background

❖ Weakly Supervised approaches

- paired data에 대한 의존도를 줄이기 위해, weakly supervised methods는 **pseudo-captions**를 활용한다.
- 그러나, pseudo-captions는 이미지와 관련 없는 단어들을 포함한다.
- 또한, pseudo-captions 생성을 위해서는 **고정된 labeled data로 pre-train된 object detector**가 필요하다.

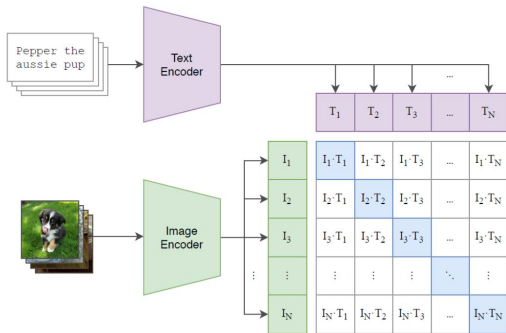
❖ Unsupervised approaches

• CLIP

: 거대 web-collected paired data 활용한 V-L pre-train 모델

<CLIP의 pre-train 과정>

- 1) N개의 이미지, 텍스트를 각각 인코더에 통과시켜 임베딩 벡터 산출
- 2) 두 벡터의 내적을 통해 image-text 코사인 유사도 계산
- 3) **positive pair**에서의 코사인 유사도는 최대화하고, **negative pair**에서의 유사도는 최소화하도록 CE loss를 사용하여 V-L 유사도 학습



→ CLIP의 'Image, Text 활용한 pre-train'은 Vision-Language task에서의 zero-shot 가능성을 보여준다.

Background

❖ Unsupervised approaches

- 대표적인 zero-shot image captioning : **Zerocap**

- ⇒ CLIP으로부터 학습된 V-L knowledge를 활용하여 계산된 clip score로 image-text 매칭을 유도하고,
- ⇒ GPT-2로부터 학습된 linguistic knowledge를 활용하여 자연스러운 캡션 생성하게끔 한다.
- ⇒ 두 모델의 재학습이나 fine-tuning 없이도 추론이 가능 (zero-shot)

- 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)
- ⇒ 추론에 많은 시간 소요, 실생활에서의 적용 어려움

→ MAGIC은 많은 연산을 피하기 위해, **MAGIC Search**를 활용하여 visual 정보를 decoding 과정에 directly하게 적용한다.

Background

❖ Decoding Strategy in NLP

- GPT와 같은 디코더 모델에서 취하는 기존 디코딩 전략은 크게 두 가지 방법론으로 나누어볼 수 있다.

1. Deterministic Methods

: 언어모델에 의해 측정된 likelihood가 가장 높은 텍스트 토큰을 연속적으로 선택하여 생성하는 방법

$$w_t = \operatorname{argmax}_w P(w|w_{1:t-1})$$

1) Greedy Search

- 매 time step 마다 확률 분포 상 가장 높은 확률을 갖는 토큰을 next token으로 선택한다.

2) Beam Search

- Greedy search와 다르게, 매 step마다 탐색의 영역을 k개의 가장 likelihood가 높은 토큰들로 유지하며 다음 단계를 탐색한다.

- 시퀀스 길이 N만큼의 시점 t가 존재하기 때문에, 한 번이라도 정답이 아닌 다른 토큰으로 예측하게 되면 뒤의 디코딩에도 영향을 미치게 되고, 불필요한 단어 반복 등 model degeneration 문제가 발생할 수 있다.

Background

❖ Decoding Strategy in NLP

2. Stochastic Methods

: 다음에 올 토큰에 대한 확률분포에 따라 단어를 샘플링하는 방식

3) Top-k Sampling

$\sum_{v \in V^{(k)}} p_{\theta}(v|\mathbf{x})$ 을 최대로 하는 단어 집합 V 로부터 Sampling을 수행한다.

4) Nucleus Sampling

$\sum_{v \in U} p_{\theta}(v|\mathbf{x}) \geq p$ 누적 확률이 확률 p 에 가까운 최소한의 단어 집합 U 로부터 Sampling을 한다.

→ Model degeneration 문제는 완화했지만, 텍스트 의미를 유지하는데 한계가 있다. (Sementic Inconsistency)

Background

❖ Decoding Strategy in NLP

- Contrastive Search

: degeneration penalty를 추가하여, 문법적, 의미적으로 모두 자연스러운 text를 생성한다.

$$x_t = \arg \max_{v \in V^{(k)}} \left\{ (1 - \alpha) \times \underbrace{p_{\theta}(v | \mathbf{x}_{<t})}_{\text{model confidence}} - \alpha \times \underbrace{(\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})}_{\text{degeneration penalty}} \right\}$$

$\alpha \in [0, 1]$: 두 요소의 가중치 조절 하이퍼파라미터. ($\alpha = 0$ 일 때 greedy search로 degenerate 됨)

모델의 분포로부터 top-k 예측 단어들을 후보 집합으로 정의해,

(1) model confidence와 (2) degeneration penalty 를 모두 고려한 값이 최대가 되도록 단어를 Sampling 한다.



(1) 가장 확률 높은 후보 단어를 선택해 prefix text 와의 의미적 일관성을 유지하면서도,

(2) 두 토큰의 코사인 유사도를 반영하여 이전 context와 구별되는 의미있는 단어를 선택 할 수 있다.

Background

→ MAGIC은 앞서 설명한 Contrastive Search에 MAGIC score를 추가하여, CLIP의 이미지 정보를 활용한 디코딩을 수행한다.

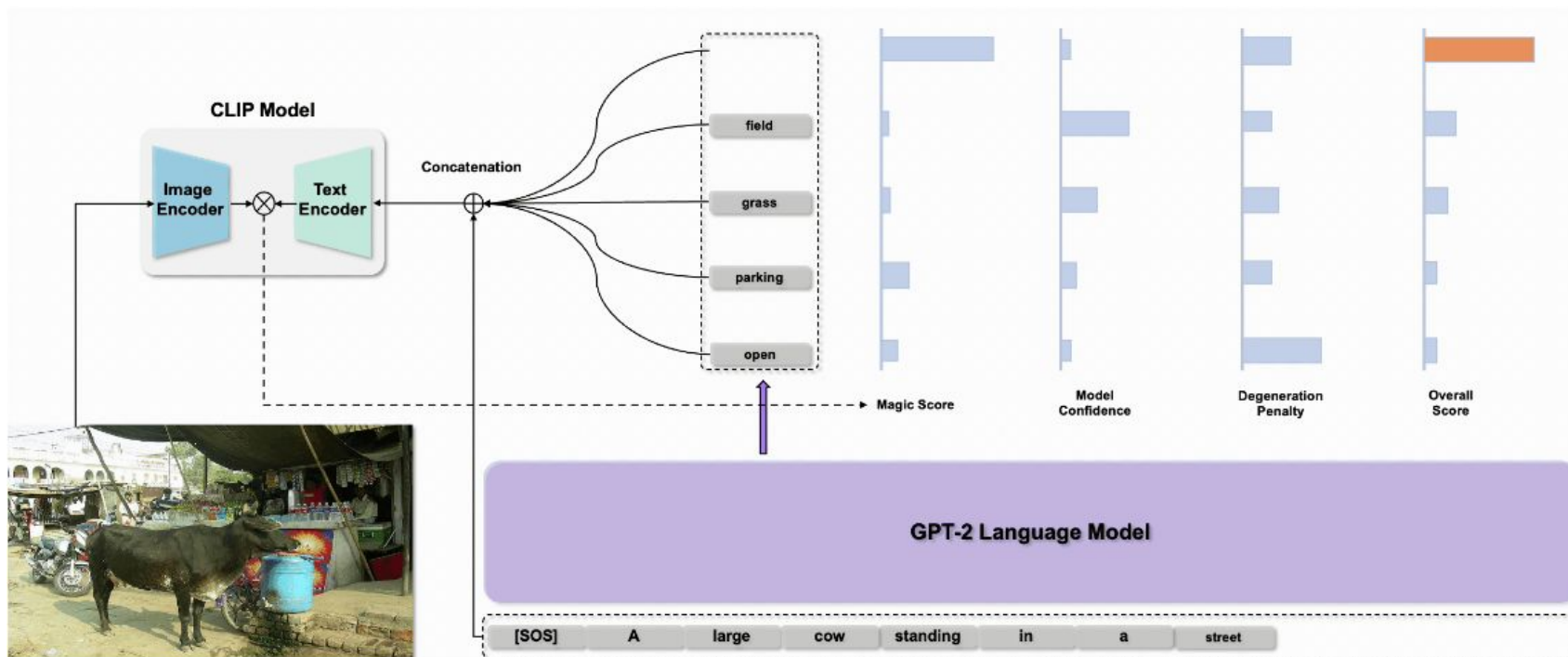
❖ “Plug and Play”

- MAGIC이 visual 정보를 decoding 과정에 directly하게 적용할 수 있는 것은 “Plug and Play” 방법으로 볼 수 있다.
- “Plug and Play” 방법 활용한 기존 연구
 - PPGN -> Replaceable condition network 활용하여 조건에 따른 sampling이 가능하다.
 - PPLM -> Attribute Model을 추가하여 다음 단어의 예측을 원하는대로 conditional하게 조정한다.

→ MAGIC은 기존과 다르게 언어생성 모델의 decoding strategy에 visual-related term을 추가하여 "plug and play"

MAGIC (iMAge-Guided text generation with CLIP)

MAGIC의 전반적인 아키텍처는 다음과 같다.

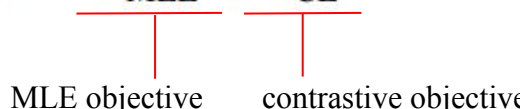


MAGIC (iMAge-Guided text generation with CLIP)

Methodology

1) Unsupervised Language Modelling

- 언어생성모델은(ex. GPT-2) 의미있는 text 생성 가능하지만, fine-tuning 없이 text decoding 진행 시 특정 세부 task에 관한 text가 부자연스러운 경우가 존재한다.
- 따라서 본 논문에서는 **decoding 하기 전, task 별 text corpus로 언어모델 fine tuning을 수행한다.**
(decoding 과정에서는 fine-tuning 된 GPT-2와 CLIP은 고정시키고 추가학습 하지 않음)
- 언어모델의 파라미터 θ 를 최적화하기 위한 목적함수 정의

$$: \quad \mathcal{L} = \mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{CL}}$$


MLE objective contrastive objective

MAGIC (iMAge-Guided text generation with CLIP)

Methodology

1) Unsupervised Language Modelling

- 먼저, 언어모델의 파라미터 θ 를 학습시키기 위한 목적함수 **Maximum Likelihood Estimation (MLE)**는 다음과 같다.

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|\mathbf{x}|} \sum_{i=1}^{|\mathbf{x}|} \log p_{\theta}(x_i | \mathbf{x}_{<i}).$$

- 추가적으로, 토큰끼리의 코사인 유사도를 반영한 **contrastive objective**(\mathcal{L}_{CL}) 을 다음과 같이 정의한다.

$$\mathcal{L}_{\text{CL}} = \frac{1}{|\mathbf{x}| \times (|\mathbf{x}| - 1)} \sum_{i=1}^{|\mathbf{x}|} \sum_{j=1, j \neq i}^{|\mathbf{x}|} \max\{0, \overbrace{\rho}^{\text{표현공간 분포 정규화를 위한 pre-defined margin, } [-1,1]} - s(h_{x_i}, h_{x_i}) + s(h_{x_i}, h_{x_j})\} \quad \text{cosine similarity}$$

- Contrastive objective**는 Cosine similarity 통해 구별된 토큰의 분포가 더욱 discriminative 하도록 모델의 표현공간을 보정한다.

=> 모델의 Generality를 향상시키고, 다양한 task에 적용 가능하게 한다. (ex. Story Generation..)

MAGIC (iMAge-Guided text generation with CLIP)

Methodology

1) Unsupervised Language Modelling

$$\textcircled{1} \mathcal{L}_{\text{MLE}} = -\frac{1}{|\mathbf{x}|} \sum_{i=1}^{|\mathbf{x}|} \log p_{\theta}(x_i | \mathbf{x}_{<i}).$$

$$\textcircled{2} \mathcal{L}_{\text{CL}} = \frac{1}{|\mathbf{x}| \times (|\mathbf{x}| - 1)} \sum_{i=1}^{|\mathbf{x}|} \sum_{j=1, j \neq i}^{|\mathbf{x}|} \max\{0, \rho - s(h_{x_i}, h_{x_i}) + s(h_{x_i}, h_{x_j})\}.$$

앞서 설명한 두 가지 objective를 결합한 최종적인 LM 목적함수는 다음과 같다.

$$\mathcal{L} = \mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{CL}}$$

MLE objective

contrastive objective

MAGIC (iMAge-Guided text generation with CLIP)

Methodology

2) MAGIC Search

논문에서 제안한 decoding strategy "MAGIC Search"는 visual 정보에 따라 decoding을 가이드한다.

매 time step t 마다 생성되는 output token x_t 는 다음과 같이 표현할 수 있다.

$$x_t = \arg \max_{v \in V^{(k)}} \left\{ (1 - \alpha) \times \underbrace{p_{\theta}(v | \mathbf{x}_{<t})}_{\text{model confidence}} - \underbrace{\alpha \times (\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})}_{\text{degeneration penalty}} + \beta \times \underbrace{f(v | \mathcal{I}, \mathbf{x}_{<t}, V^{(k)})}_{\text{magic score}} \right\}$$

- 표시된 항은 top-k predictions 집합 V_k 에 속한 token에 대해 degeneration penalty를 반영한 예측 확률을 나타낸다.
- model confidence에 **degeneration penalty**를 추가한 것은 앞서 설명한 **contrastive search**에서 소개한 decoding 방법이다.
- 코사인 유사도를 반영한 degeneration penalty를 적용하여 모델의 degeneration을 방지함

MAGIC (iMAge-Guided text generation with CLIP)

Methodology

2) MAGIC Search

디코딩 과정에 visual control을 적용하기 위해, **magic score**를 추가한다.

$$x_t = \arg \max_{v \in V^{(k)}} \left\{ (1 - \alpha) \times \underbrace{p_{\theta}(v | \mathbf{x}_{<t})}_{\text{model confidence}} - \underbrace{\alpha \times (\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})}_{\text{degeneration penalty}} + \beta \times \underbrace{f(v | \mathcal{I}, \mathbf{x}_{<t}, V^{(k)})}_{\text{magic score}} \right\}$$

$$f(v | \mathcal{I}, \mathbf{x}_{<t}, V^{(k)}) = \frac{e^{\text{CLIP}(\mathcal{I}, [\mathbf{x}_{<t}:v])}}{\sum_{z \in V^{(k)}} e^{\text{CLIP}(\mathcal{I}, [\mathbf{x}_{<t}:z])}} = \frac{e^{h_{\mathcal{I}}^{\top} h_{[\mathbf{x}_{<t}:v]}}}{\sum_{z \in V^{(k)}} e^{h_{\mathcal{I}}^{\top} h_{[\mathbf{x}_{<t}:z]}}}$$

- magic score 는 후보집합 V_k 에 속한 토큰들에 대한 CLIP기반 image-text 유사도를 나타낸다.
 - 이미지 정보를 반영하여 의미적으로 관련된 next token을 예측할 수 있도록 한다.
 - $\beta = 0 \rightarrow$ visual control 반영 안 됨 (기본적인 contrastive search로 degenerate)

MAGIC (iMAge-Guided text generation with CLIP)

Methodology

- 1) 새롭게 정의한 LM의 목적함수와

$$\mathcal{L} = \mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{CL}}$$

- 2) MAGIC score를 추가한 디코딩 방법은 (MAGIC Search)

$$x_t = \arg \max_{v \in V^{(k)}} \left\{ (1 - \alpha) \times \underbrace{p_{\theta}(v | \mathbf{x}_{<t})}_{\text{model confidence}} - \underbrace{\alpha \times (\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})}_{\text{degeneration penalty}} + \beta \times \underbrace{f(v | \mathcal{I}, \mathbf{x}_{<t}, V^{(k)})}_{\text{magic score}} \right\}$$

→ 추가적인 지도 학습이나 gradient update 없이 visual control을 디코딩 과정에 directly plugging 할 수 있게 한다.

Experiments

❖ Zero-Shot Image Captioning

먼저, Zero-Shot Image Captioning에 대한 실험을 진행

● Fine-tuning GPT-2

- benchmark data set (MS-COCO, Flickr30k) 으로 각각 GPT-2 fine-tuning 진행한다.
 - **MS-COCO** : k , α , and $\beta = 45, 0.1, 2.0$
 - **Flickr30k** : k , α , and $\beta = 25, 0.1, 2.0$
 - Optimizer : Adam (learning rate : $2e-5$), epoch : 3, contrastive loss margin $p : 0.5$
- => validation set 성능에 근거한 MAGIC Search 파라미터

● Baseline zero-shot methods

1) 기존 decoding methods

- top-k sampling ($k=40$), nucleus sampling ($p=0.95$), contrastive search ($\alpha, \beta =$ magic search에 적용한 값과 동일)
(위 세 가지 methods는 디코딩 과정에서 image 정보 고려하지 않음)

2) **CLIPRe** : CLIP 유사도를 통해 가장 관련성이 높은 캡션을 검색

3) **ZeroCap**

+ 기존 Supervised Methods

Experiments

Results

Model	MS-COCO						Flickr30k						Speed
	B@1	B@4	M	R-L	CIDEr	SPICE	B@1	B@4	M	R-L	CIDEr	SPICE	
Supervised Approach													
BUTD	77.2	36.2	27.0	56.4	113.5	20.3	-	27.3	21.7	-	56.6	16.0	-
GVD	-	-	-	-	-	-	66.9	27.3	22.5	-	62.3	16.5	-
UniVLP	-	36.5	28.4	-	116.9	21.2	-	30.1	23.0	-	67.4	17.0	-
ClipCap	-	33.5	27.5	-	113.1	21.1	-	-	-	-	-	-	-
Oscar	-	36.5	30.3	-	123.7	23.1	-	-	-	-	-	-	-
LEMON	-	40.3	30.2	-	133.3	23.3	-	-	-	-	-	-	-
Weakly Supervised Approach													
UIC	41.0	5.6	12.4	28.7	28.6	8.1	-	-	-	-	-	-	-
IC-SME	-	6.5	12.9	35.1	22.7	-	-	7.9	13.0	32.8	9.9	-	-
S2S-SS	49.5	6.3	14.0	34.5	31.9	8.6	-	-	-	-	-	-	-
S2S-GCC	50.4	7.6	13.5	37.3	31.8	8.4	-	-	-	-	-	-	-
Unsupervised Approach													
Top- <i>k</i>	33.6	2.4	8.3	25.6	3.8	1.7	34.0	2.9	9.0	24.4	3.3	2.7	69.9×
Nucleus	32.6	2.3	7.8	24.8	3.1	1.4	32.6	2.4	8.1	23.4	2.5	2.4	72.5 ×
Contrastive	39.5	3.0	10.8	30.8	7.7	2.9	37.6	4.3	9.8	25.7	8.9	4.6	50.4×
CLIPRe	39.5	4.9	11.4	29.0	13.6	5.3	38.5	5.2	11.6	27.6	10.0	5.7	-
ZeroCap	49.8	7.0	15.4	31.8	34.5	9.2	44.7	5.4	11.8	27.3	16.8	6.2	1.0×
MAGIC	56.8	12.9	17.4	39.9	49.3	11.3	44.5	6.4	13.1	31.6	20.4	7.1	26.6×

Table 1: Image Captioning Results on MS-COCO and Flickr30k.

Experiments

❖ Zero-Shot Image Captioning

Results

Unsupervised Approach														
Top- <i>k</i>	33.6	2.4	8.3	25.6	3.8	1.7	34.0	2.9	9.0	24.4	3.3	2.7	69.9×	
Nucleus	32.6	2.3	7.8	24.8	3.1	1.4	32.6	2.4	8.1	23.4	2.5	2.4	72.5 ×	
Contrastive	39.5	3.0	10.8	30.8	7.7	2.9	37.6	4.3	9.8	25.7	8.9	4.6	50.4×	
CLIPRe	39.5	4.9	11.4	29.0	13.6	5.3	38.5	5.2	11.6	27.6	10.0	5.7	-	
ZeroCap	49.8	7.0	15.4	31.8	34.5	9.2	44.7	5.4	11.8	27.3	16.8	6.2	1.0×	
MAGIC	56.8	12.9	17.4	39.9	49.3	11.3	44.5	6.4	13.1	31.6	20.4	7.1	26.6×	

- image input 에 대한 조건 없이 language model 만 사용했을 시 (Top-k, Nucleus, Contrastive), 의미있는 캡션 생성이 어려운 것을 보여준다.
- CLIPRe : training 과 test sets의 데이터 불일치로 인한 격차로 인해 ZeroCap 보다 좋지 못한 성능을 보인다.
- MAGIC
 - 11개의 metrics에서 가장 높은 성능 달성함
 - Zerocap보다 27배 빠른 속도를 보인다. (실용적 사용 가능성 보임)
=> gradient updates 포함하지 않기 때문

Experiments

❖ Zero-Shot Image Captioning

추가적으로, 모델의 일반화 성능을 측정하기 위해 **Cross-Domain Experiment**을 수행했다.

- 1) **MS-COCO**의 text corpus로 Fine-tuning 한 모델을 **Flickr30k** 데이터 셋으로 추론
- 2) **Flickr30k**의 text corpus로 Fine-tuning 한 모델을 **MS-COCO** 데이터 셋으로 추론

Results

Model	MS-COCO \Rightarrow Flickr30k						Flickr30k \Rightarrow MS-COCO					
	B@1	B@4	M	R-L	CIDEr	SPICE	B@1	B@4	M	R-L	CIDEr	SPICE
Top- <i>k</i>	34.9	2.4	7.5	24.2	2.3	1.7	30.0	1.8	8.5	23.6	2.5	1.7
Nucleus	33.4	1.7	7.0	23.3	1.8	1.3	29.1	1.6	8.0	22.9	2.1	1.6
Contrastive	40.3	5.3	10.7	30.5	5.1	3.4	33.8	3.2	10.2	25.5	4.2	3.7
CLIPRe	38.7	4.4	9.6	27.2	5.9	4.2	31.1	3.0	9.9	22.8	8.5	3.9
MAGIC	46.4	6.2	12.2	31.3	17.5	5.9	41.4	5.2	12.5	30.7	18.3	5.7

Table 2: Cross-Domain Evaluation. $X \Rightarrow Y$ means source domain \Rightarrow target domain.

- 모든 methods가 In-domain results 보다 낮은 성능을 보였지만, **MAGIC**은 다른 methods에 비해 좋은 성능을 보인다.
 \Rightarrow **Robustness & Generalization** 성능 보여줌

Experiments

❖ Zero-Shot Image Captioning

Qualitative Evaluation



Reference	Many different signs cover a post next to a bus stop.
CLIPRe	A row of men using laptops on side of a building.
ZeroCap	A school bus on a sign street.
MAGIC	A street sign with a building in the background.

(a)



Reference	Two kayaks, one pink the other yellow, on bank of water.
CLIPRe	A kitchen has wood cabinets, a dishwasher, sink, and refrigerator.
ZeroCap	a boatboard and a small boat in a small boat.
MAGIC	A yellow boat is lined up on the beach.

(d)

CLIPRe : 관련 없는 단어 "building"을 포함

ZeroCap : 이미지에 없는 단어 "school bus"를 설명

MAGIC : "street sign" 정확히 표현

CLIPRe : 이미지와 무관

ZeroCap : 관련된 물체 표현했으나, 문법적으로 자연스럽지 않음

MAGIC : 관련된 물체 표현하면서, 문법적으로 자연스러운 문장 생성

Experiments

❖ Story Generation

모델의 범용성과 확장성을 확인하기 위해, Story Generation에 대한 실험도 진행하였다.

- **Story Generation** : story title (i.e., text prompt)을 주었을 때, 언어 모델이 관련된 스토리를 생성하게끔 함
- **Model and Baselines**
 - ROCStories benchmark로 Fine-tuning 된 GPT-based Language Model 사용
 - **Baselines**
 - (1) Greedy search
 - (2) Beam search (beam width = 10)
 - (3) Top-k sampling ($k = 40$)
 - (4) Nucleus sampling ($p = 0.95$)
 - (5) Typical sampling ($\tau = 0.2$)
 - (6) Contrastive search ($k = 5, \alpha = 0.6$)

=> 각 methods의 하이퍼파라미터는 val set의 optimal MAUVE 점수에 근거하여 결정

Experiments

❖ Story Generation

- Implementation Details of MAGIC

: MAGIC은 Story Title로 text prompt 대신 **image input**을 사용한다.

- 1) CLIP으로 계산된 score 활용하여 **Story title**과 가장 관련있는 **Image**를 찾는다. (Conceptual Caption Dataset)
- 2) 찾은 이미지로부터 관련된 **story text**를 생성한다. (by using MAGIC Search, $k = 5$, $\alpha = 0.6$, $\beta = 0.15$)

Results

Method	Automatic Evaluation							Human Evaluation			
	rep-2↓	rep-3↓	rep-4↓	div.↑	coh.↑	MAUVE↑	CLIPScore↑	coh.↑	flu.↑	inform.↑	si-rel.↑
Agreement	-	-	-	-	-	-	-	0.68	0.57	0.66	0.73
Greedy	22.27	15.42	12.36	0.58	0.473	0.53	0.23	2.67	3.20	3.10	2.03
Beam	26.76	21.79	18.85	0.47	0.478	0.46	0.25	2.71	3.23	3.15	2.05
Top- k	3.38	0.76	0.23	0.95	0.458	0.86	0.21	2.52	3.69	3.62	1.96
Nucleus	2.92	0.60	0.18	0.96	0.452	0.88	0.21	2.48	3.68	3.71	1.92
Typical	2.52	0.46	0.12	0.97	0.450	0.84	0.19	2.32	3.70	<u>3.76</u>	1.75
Contrastive	2.49	0.38	0.09	0.97	<u>0.488</u>	<u>0.89</u>	<u>0.28</u>	<u>2.86</u>	<u>3.72</u>	<u>3.76</u>	<u>2.07</u>
MAGIC	<u>2.51</u>	0.38	0.09	0.97	0.514	0.91	0.36	3.20★	3.76	3.85	2.40★
Human	2.21	0.37	0.09	0.97	0.542	1.00	0.40	3.77	4.11	4.22	2.59

Experiments

❖ Story Generation

Results

Automatic Evaluation

- MAGIC이 모든 Evaluate metrics에서 **가장 좋은 성능** 보임
=> MAGIC이 다양하면서도 human-written stories와 가장 유사한 stories 생성한다.
- 높은 semantic coherence - Story title로 찾은 이미지가 다양한 관련된 visual 특징을 포함하고 있다.
 - MAGIC은 이러한 Visual 정보를 활용하여 더 많은 정보를 활용한 이미지 캡셔닝을 수행할 수 있다.

Human Evaluation

- **Fluency score** : Visual 정보 활용하여 더 다양하고 흥미로운 content 생성이 가능하다.
- **Coherence metric** : Story title 과 생성된 Story 간의 일관성이 유지된다.
- **Story-Image Relevance** : 이미지와 관련된 텍스트를 생성할 수 있다.

Experiments

❖ Story Generation

Qualitative Evaluation



	Title	Ice Cream Tasting
	Contrastive	My family went to a ice cream shop. We ordered the Ice Cream Truck. It was delicious. The customer service was terrible. We had to leave for another day.
	MAGIC	My family went to a ice cream shop. They ordered three flavors of ice cream . The first one was strawberry , the second was chocolate , and the third was orange. I was excited to try all three flavors . It was very good and I had a great time at the ice cream shop.
	Title	Sand Volleyball
	Contrastive	I went to the park yesterday. It was raining a lot. I had to use the water pump to get to the park. When I got there, there was nothing to play in the park. I ended up playing volleyball instead.
	MAGIC	I went to the beach with my friends . It was a sand volleyball game . We played for two hours. My friend got to pick his team. He won the game for his team.

Figure 2: Examples of story generation. MAGIC can generate text (highlighted in blue) that is related to the visual concepts displayed in the image. (Best viewed in color and by zooming in.)

생성된 story를 Contrastive Search 결과와 비교했을 때,

Visual guide를 하지 않은 Contrastive Search 보다 더 다양하고 자세한 정보를 포함하면서도, 일관성을 유지한다.

Conclusion

❖ Summerization

- MAGIC은 기존 image captioning 모델과 다르게, **contrastive search 디코딩 방법에 CLIP기반 magic score만을 추가**하여 "plug and play"한 방법으로 이미지 정보를 텍스트 생성 시 활용하였다.
- 이에 따라 **불필요한 연산 없이** 빠른 Image-Text task 수행이 가능한 method이다.
- **모델 아키텍처에 구애받지 않는 decoding scheme**이므로 다른 task, modal로 확장이 가능한 method이다.
- 두 가지 이미지-텍스트 task 1) Image Captioning, 2) Story Generation 에서 기존 SOTA methods 에 비해 좋은 성능을 달성했다.

❖ Contribution

- 언어 모델 fine tuning 과정이 text corpus만을 사용한 unsupervised 한 방법으로 이루어졌다는 점에서 제로샷에 가까운 모델이라고 볼 수 있다. (image-text paired data를 fine-tuning에 사용하지 않음)
- 여러 task(Story generation..)에 적용 가능하여 **generality가 높고**, ZeroCap보다 27배 빠르다는 점에서 **실용성이 높은** 모델.
(fine-tuning 없이 zero-shot만으로도 높은 성능을 달성한다면 더 좋을 것 같음)