Image Difference Captioning with Pre-training and Contrastive Learning

Data Engineering Lab

Sunghyun Ahn
skd@yonsei.ac.kr

<2023/05/17>

① Introduction

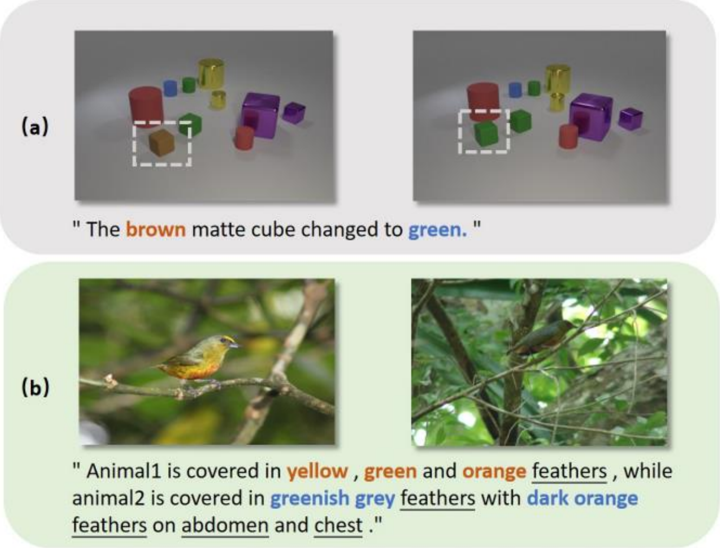
Image Difference Captioning (IDC)

- IDC는 비슷한 두 이미지 간의 시각적 차이를 자연어로 설명하는 것을 목표로 함

- **fine-grained semantic comprehension**
  - ➡ 비슷한 이미지에 대한 차이를 설명해야 됨
    (시각적 차이와 텍스트 간의 관계를 파악해야 됨)
  - ➡ 시나리오에 따라 시각적 차이가 다양할 수 있음

    (a): 기하학적 객체의 변화에 중점을 둠
    (Move, Add, Drop, Color, Texture)
    (b): 새 종류에 따른 외형의 차이에 중점을 둠

- **high-cost of manual annotation**
  - ➡ 두 이미지를 각각 관찰한 다음 차이를 비교해야 됨
  - ➡ triplet format (img1, img2, text)으로 구성해야 됨



① Introduction

New training schema for IDC (three self-supervised learning)

시각적 차이(visual differences)와 텍스트(text description) 간의 관계를 파악하고자 함

- Masked Language Modeling (MLM)

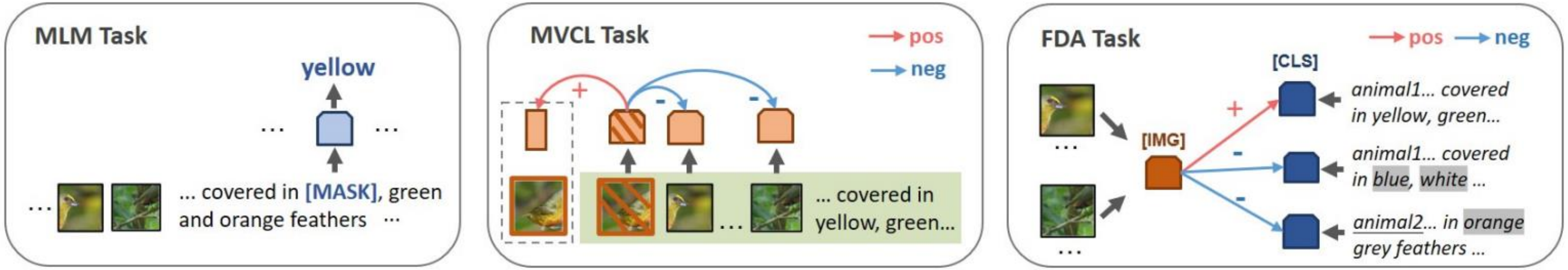
➡ 언어 모달리티에서 텍스트 토큰을 마스킹하고 해당 마스크를 복원하는 과정에서 시각-언어간의 상호 작용을 진행

- Masked Contrastive Learning (MVCL)

➡ 시각 모달리티에서 이미지 패치를 마스킹하고 해당 마스크를 복원하는 과정에서 시각-언어간의 상호 작용을 진행

- Fine-grained Difference Aligning (FDA)

➡ Hard Negative Texts를 만들고 시각적 차이를 자연어와 세밀하게 대조하면서 시각-언어간의 상호 작용을 진행



Introduction

Data expansion strategy

추가적인 cross-task 데이터로부터 배경 지식을 학습하고자 함 (cross-task data: 서로 다른 작업에서 수집된 데이터)

General Image Captioning (GIC)

→ 이미지와 텍스트 간의 대응을 학습하면서 이미지의 시각적 특징과 텍스트의 언어적 특징의 관계를 이해함

Fine Grained Visual Classification (FGVC)

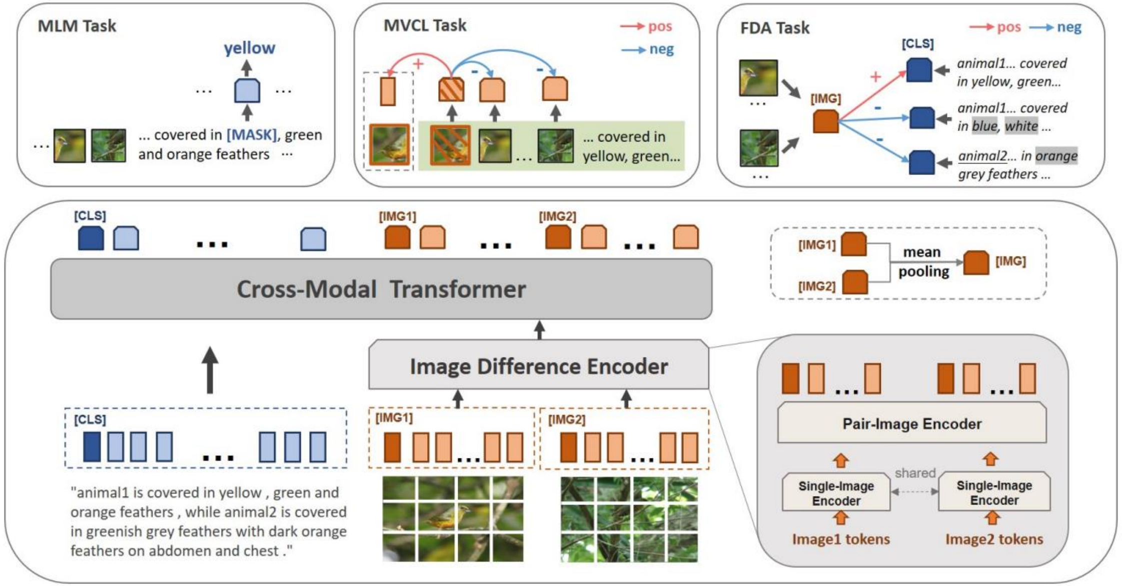
→ 비슷한 종류의 물체들 사이의 미묘한 차이를 인식하고 분류하는 작업으로, 더 세부적인 시각적 표현을 학습함

② Method

Proposed Model

Image Difference Encoder를 통해 세밀한 이미지 차이를 파악함

Cross-Modal Transformer를 통해 시각적 정보와 텍스트 정보 간 관계를 파악함



$$\{V^{(1)}, V^{(2)}, T\}$$

$$V^{(1)} = \{[\mathrm{IMG1}], v_0^{(1)}, \ldots, v_i^{(1)}, \ldots, v_N^{(1)}\}$$

$$V^{(2)} = \{[\mathrm{IMG2}], v_0^{(2)}, \ldots, v_i^{(2)}, \ldots, v_N^{(2)}\}$$

$$T = \{[\mathrm{CLS}], [\mathrm{BOS}], w_0, \ldots, w_M, [\mathrm{EOS}]\}$$

$$\widetilde{V}^{(1)}, \widetilde{V}^{(2)} = \mathcal{F}_{\text{pair}}\left(\mathcal{F}_{\text{sing}}(V^{(1)}), \mathcal{F}_{\text{sing}}(V^{(2)})\right)$$

$$\hat{V}^{(1)}, \hat{V}^{(2)}, \hat{T} = \mathcal{F}_{\text{cross}}\left(\widetilde{V}^{(1)}, \widetilde{V}^{(2)}, T\right)$$

Method

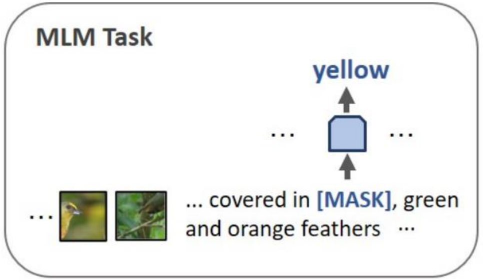
Objective Function

Masked Language Modeling (MLM) ➡ mask 15%

$$\mathcal{L}_{\mathrm{MLM}} = \mathbb{E}_{V,T \in D}\left[-\log P_\theta\left(w_m \mid w_{\setminus m}, \widetilde{V}^{(1)}, \widetilde{V}^{(2)}\right)\right]$$

w_m: masked word

$w_{\setminus m}$: unmasked word



Masked Contrastive Learning (MVCL) ➡ mask 15%

$$\mathcal{L}_{\mathrm{MVCL}} = \mathbb{E}_{V,T \in D} f_\theta\left(v_m \mid v_{\setminus m}, T\right)$$

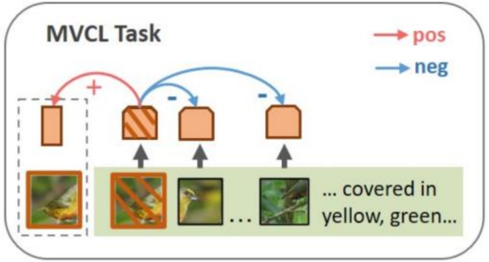
$$-\log \frac{\exp\left(d(v_m, v_m^+)/\tau_1\right)}{\exp\left(d(v_m, v_m^+)/\tau_1\right) + \sum_{v' \in \mathcal{N}(v_m)} \exp\left(d(v_m, v')/\tau_1\right)}$$

d : cosine similarity

v_m : masked image feature

v^+_m: original image feature of v_m

$\mathcal{N}(v_m)$: unmasked image features in the batch



② Method

Objective Function

Fine-grained Difference Aligning (FDA)

Retrieve: TF-IDF 유사도를 통해 검색된 비슷한 문장

Replace: 명사, 형용사 등을 다른 단어로 대체한 문장

Confuse: 주어의 위치를 서로 변경한 문장

$$\mathbb{E}_{V,T\in D}\left[-\log \mathrm{NCE}(V,T)\right],$$

$$\frac{\exp\left(d\left(V,T^{+}\right)/\tau_2\right)}{\exp\left(d\left(V,T^{+}\right)/\tau_2\right)+\sum_{T^{-}\in\mathcal{N}_T}\exp\left(d\left(V,T^{-}\right)/\tau_2\right)}$$

V : the average of special token [IMG1] and [IMG2]

T^{+} : Positive Text (Ground Truth)

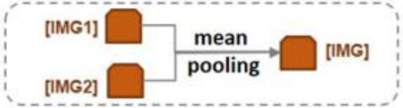
T^{-} : Negative Text (Retrieve, Replace, Confuse)

Original animal1 is brown with white tuft while animal2 is orange

Retrieve animal1 is brown with white tuft while animal2 is dark brown with grey tuft

Replace selected words [tuft, orange, brown]
animal1 is stocky with white spotting while animal2 is greenish

Confuse animal2 is brown with white tuft while animal1 is orange



② Method

Training Method with cross-task data

Pre-training

(1) GIC 데이터를 통해 학습하여 모델 파라미터를 초기화함

L_{MLM} , L_{MVCL} , L_{FDA}

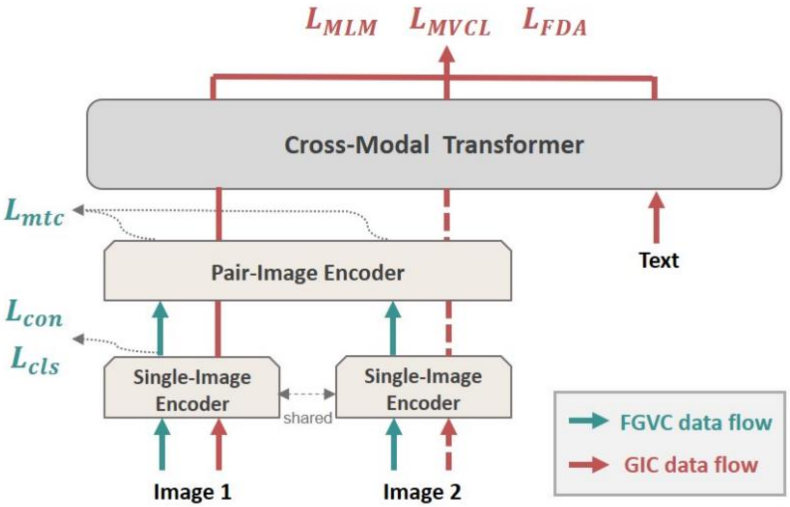
(2) FGVC 및 IDC 데이터를 통해 학습함

$L_{contrastive}$, $L_{classification}$, $L_{matching}$, L_{MLM} , L_{MVCL} , L_{FDA}

$L_{contrastive}$: batch내에서 같은 라벨은 유사도를 높이고 다르면 낮추도록 학습

$L_{classification}$: 이미지 분류를 잘 수행하도록 학습

$L_{matching}$: 두 이미지가 같은 라벨인지 아닌지를 맞추는 학습



Finetuning

IDC 데이터를 통해 전체 모델을 파인튜닝함 [L_{MLM}]

Uni-directional attention mask를 사용함

3 Experiments

Results

Birds-to-Words (GIC: CUB, FGVC: NABirds)

Model	B4	M	C(D)	**R**
Neural Naturalist (2019)	22.0	-	25.0	43.0
Relational Speaker (2019)	21.5	22.4	5.8	43.4
DUDA (2019)	23.9	21.9	4.6	44.3
L2C (2021)	31.3	-	15.1	45.3
L2C(+CUB) (2021)	**31.8**	-	16.3	45.6
Ours	28.0	23.1	18.6	48.4
Ours(+Extra Data)	31.0	**23.4**	**25.3**	**49.1**

Table 1: Comparison with state-of-the-art models on **Birds-to-Words** dataset. B4, M, R, and C(D) are short for BLEU-4, METEOR, ROUGE-L and CIDEr(D). The main metric ROUGE-L on this dataset is highlighted.

CLEVER-Change

Model	B4	M	R	**C**
Capt-Dual-Att (2019)	43.5	32.7	-	108.5
DUDA (2019)	47.3	33.9	-	112.0
VAM (2020)	50.3	37.0	69.7	114.9
VAM+ (2020)	**51.3**	**37.8**	70.4	115.8
IFDC (2021a)	49.2	32.5	69.1	118.7
DUDA+Aux (2021)	51.2	37.7	70.5	115.4
Ours	51.2	36.2	**71.7**	**128.9**

Table 2: Comparison with state-of-the-art models on **CLEVR-Change** dataset. B4, M, R, and C are short for BLEU-4, METEOR, ROUGE-L and CIDEr. The main metric CIDEr on this dataset is highlighted.

3 Experiments

Results

CIDEr performance on different type

Model	C	T	M	A	D	DI
DUDA	120.4	86.7	56.4	108.2	103.4	110.8
VAM+	122.1	98.7	82.0	126.3	115.8	122.6
IFDC	**133.2**	99.1	**82.1**	128.2	**118.5**	114.2
Ours	131.2	**101.1**	81.7	**133.3**	116.5	**145.0**

Table 3: Breakdown CIDEr performance on different type of changes of **CLEVR-Change** Dataset: C(Color), T(Texture), M(Move), A(Add), D(Drop) and DI(Distractor).

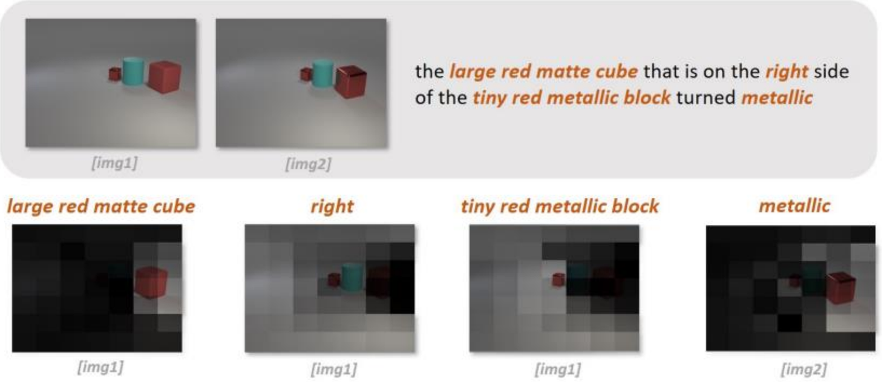


Figure 7: Visualizations of cross-modal alignment on CLEVR-Change dataset.

3 Experiments

Results

Ablation Study

Pre-training Tasks	DE	B4	M	R	C
1 None	✓	32.7	27.7	57.2	89.8
2 MLM	✓	36.7	28.2	60.9	94.9
3 MLM + MVCL	✓	50.3	37.6	70.6	119.7
4 MLM + MVCL + FDA	✓	51.2	36.2	71.7	128.9
5 MLM + MVCL + FDA	✗	49.2	35.8	68.8	107.9
6 w/o Distractor Judging	✓	49.8	36.9	69.2	123.5

Table 4: Ablation study results on CLEVR-Change dataset. **DE** is short for Image **D**ifference **E**ncoder module in our model. B4, M, R, and C are short for BLEU-4, METEOR, ROUGE-L and CIDEr. The main metric CIDEr on this dataset is highlighted.

Performance using cross-task dataset

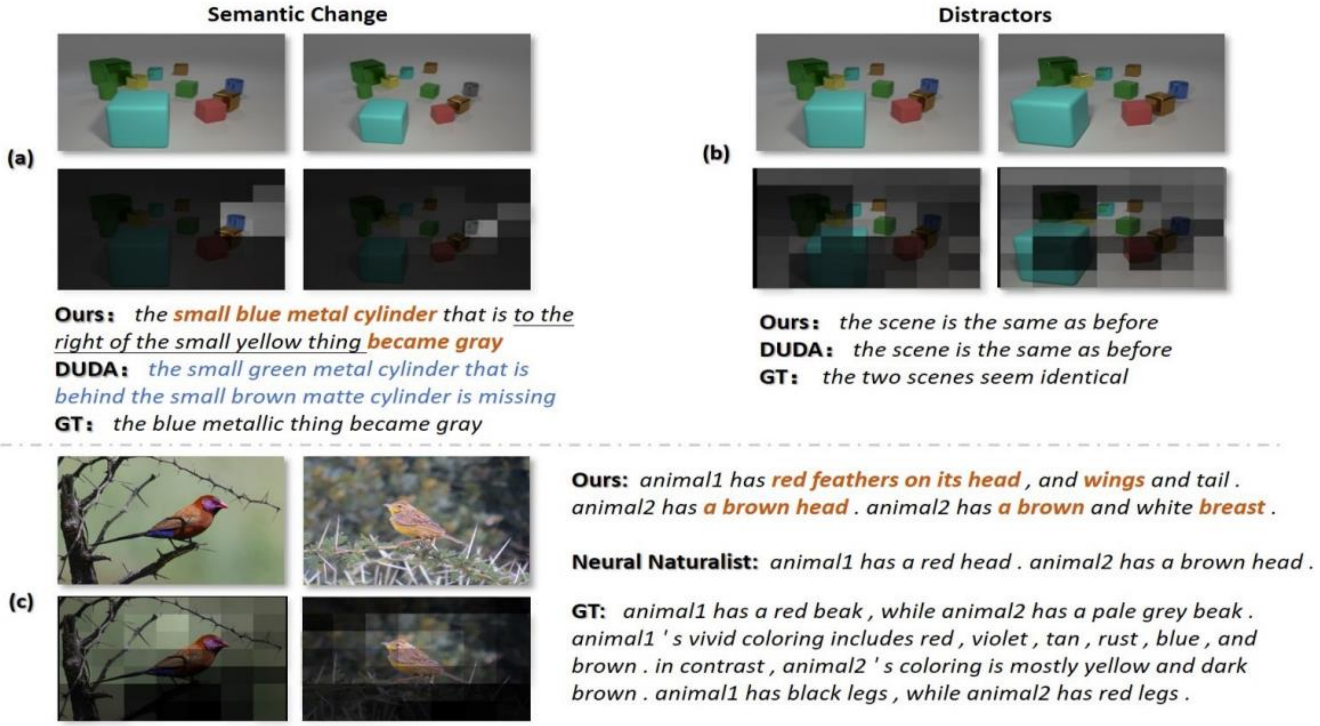
Model	B2W	CUB	NAB	B4	M	C(D)	R
L2C	✓			31.3	-	15.1	45.3
	✓	✓		31.8	-	16.3	45.6
Ours	✓			28.0	23.1	18.6	48.4
	✓	✓		29.3	23.1	23.8	48.5
	✓		✓	27.5	23.3	21.9	48.5
	✓	✓	✓	31.0	23.4	25.3	49.1

Table 5: Model performance on Birds-to-Words(B2W) dataset using two cross-task dataset including CUB and NABirds(NAB). B4, M, R, and C(D) are short for BLEU-4, METEOR, ROUGE-L and CIDEr-D. The main metric ROUGE-L on this dataset is highlighted.

3 Experiments

Results

Visualization of generated cases



4 Conclusion

Conclusion

- 세 가지 self-supervised task를 활용한 새로운 pre-training-finetuning 기법을 제안함

→ 시각적 차이(visual difference)와 텍스트(text description)간의 관계를 효과적으로 학습하였음

- 추가적인 cross-task data를 사전학습에 이용해서 부족한 IDC 데이터로 인한 한계를 극복함

→ 사전 학습으로 배경 지식을 습득하여 시각적 차이에 대한 설명을 더 잘 생성하였음

- CLEVER-Change와 Birds-to-Words 데이터셋에서 SOTA 성능을 달성함

Thank You

Data Engineering Lab

Sunghyun Ahn

skd@yonsei.ac.kr

<2023/05/17>