Improving Zero-Shot Recognition by Visual Context Embeddings

조건희

Index

Introduction

Introduction

Zero-shot recognition

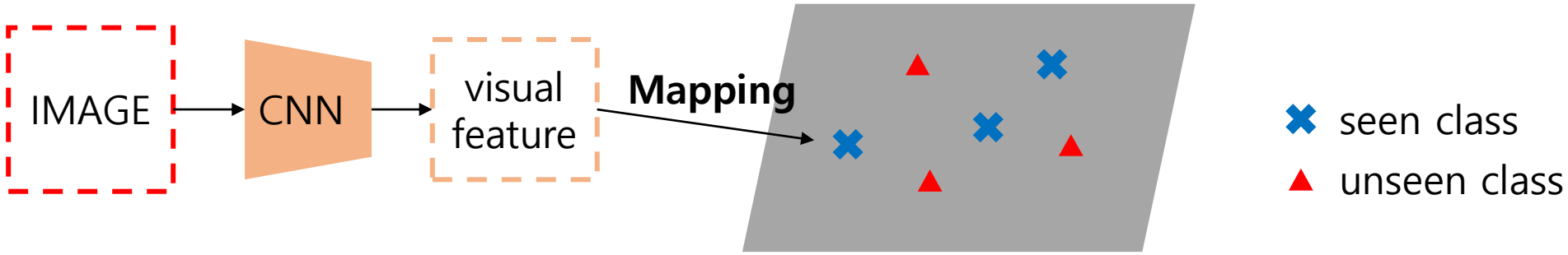
새로운 시맨틱 임베딩 스페이스를 적용하여

기존 **Zero-Shot Recognition (ZSR)**의 성능을 개선

Introduction
Zero-shot recognition

기존 Zero-Shot Recognition 의 패러다임

1) Visual feature 로부터 **시맨틱 임베딩 스페이스**로의 **매핑 함수**를 학습하면,
   - ✓ 단, 이 매핑의 학습에는 seen 클래스의 이미지와 임베딩 벡터만 사용됨.
   - ✓ 시맨틱 임베딩 스페이스에는 seen/unseen 클래스가 모두 포함되어 있어야 함.
2) 그 매핑 함수가 **Unseen 클래스의 이미지**도 알맞은 임베딩 벡터로 매핑해 줄 것이다!



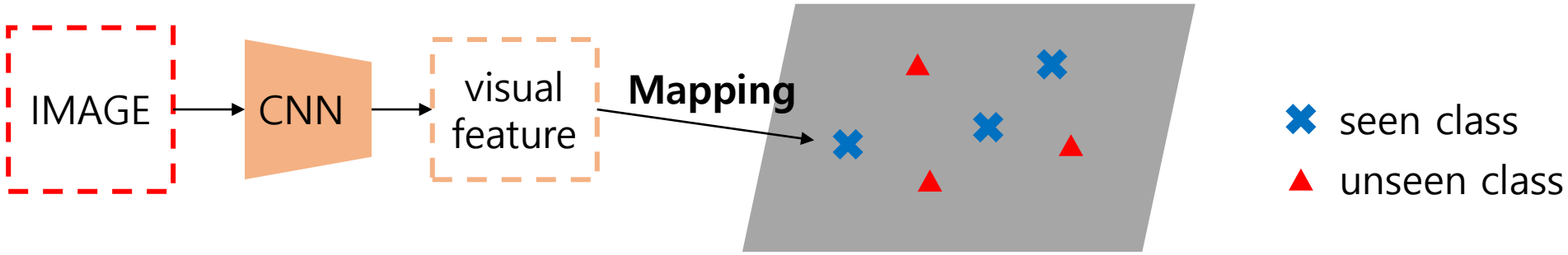
시맨틱 임베딩 스페이스

(i.e. Attribute, Knowledge graph, Word-embeddings)

Introduction
Zero-shot recognition

Assumption

- 시맨틱 임베딩 스페이스는 **seen/unseen 클래스 간의 관계정보가 함축**되어 있음.
  - 즉, 관계성이 높으면 벡터 간 유사도가 높다. (or 벡터 간 거리가 가깝다.)
- **Visual feature가 유사하면 서로 관계성이 높음.**
  - 예시 : 말/얼룩말, 고양이/호랑이, 자전거/오토바이



시맨틱 임베딩 스페이스

(i.e. Attribute, Knowledge graph, Word-embeddings)

Related work

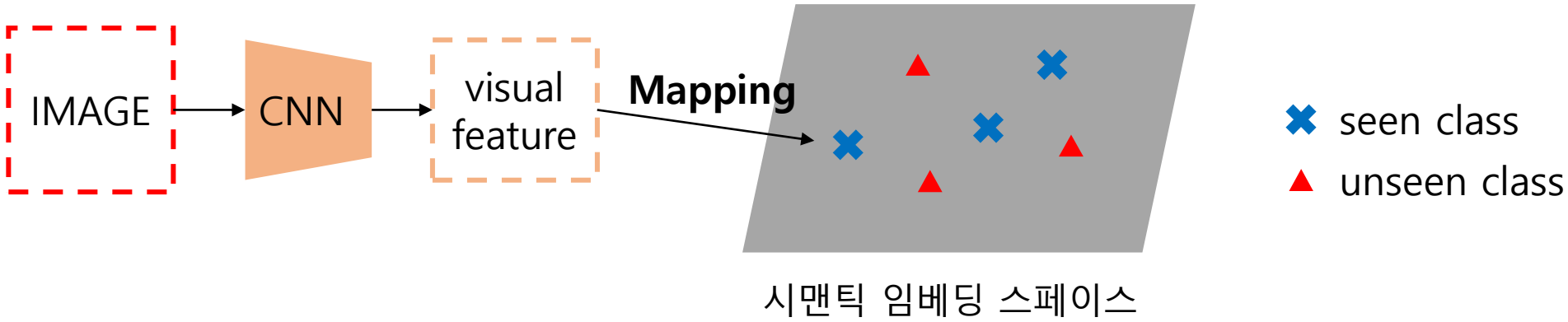
Related work

기존 ZSR 모델들

SJE, LatEm, ConSE, GCNZ

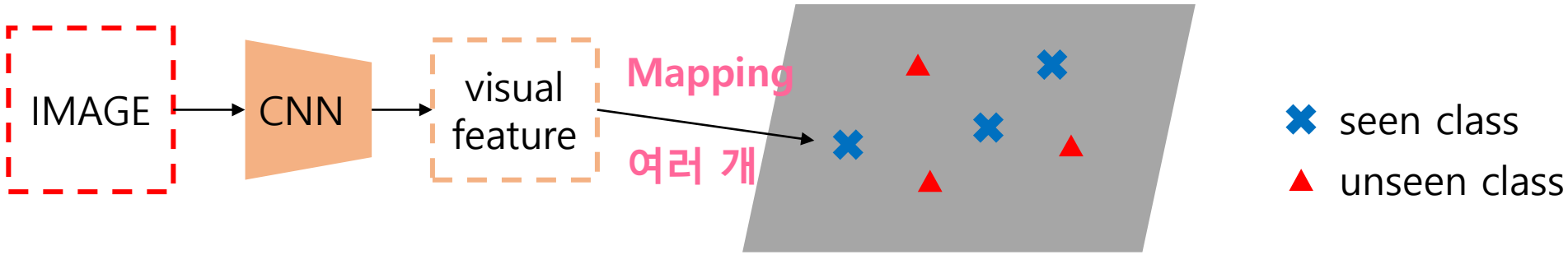
Related work

SJE



Related work

LatEm



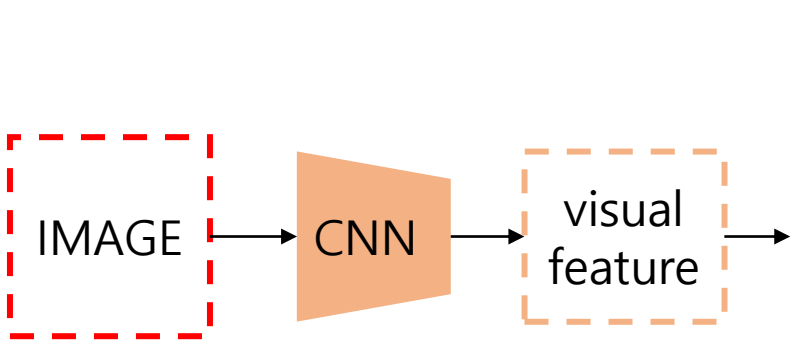
하나의 매핑 함수가 다양한 클래스의 피쳐를 전부 잘 구분하기는 어려움

→ 매핑 함수를 여러 개 사용하여 각 매핑 함수가 특정 피쳐에 대한 구별 능력을 학습할 수 있게 함.

가장 스코어가 높게 나오는 매핑 함수를 그때그때 선택적으로 사용.

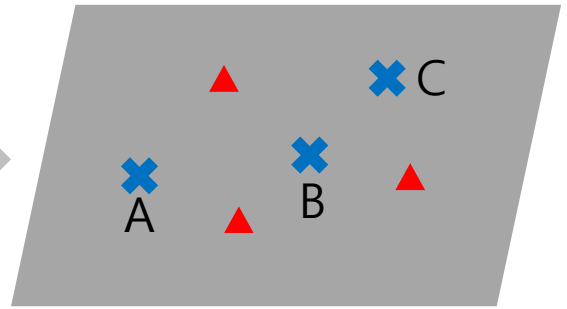
Related work

ConSE



Class	Probability
A	0.34
B	0.05
C	0.1
D	0.04
…	

Classification 결과
벡터 가중합

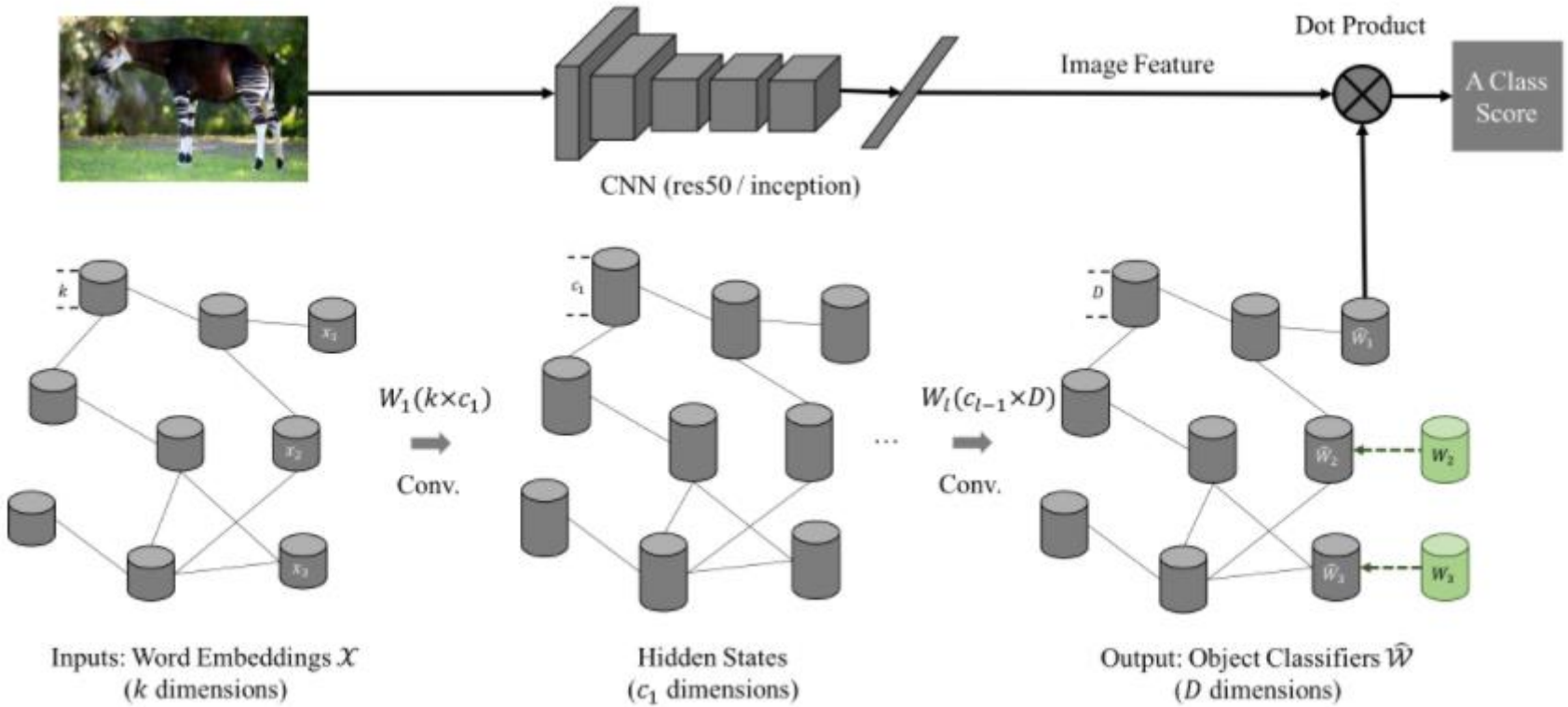


시맨틱 임베딩 스페이스

Unseen 클래스의 임베딩 벡터 = 0.34*A + 0.05*B + 0.1*C + 0.04*D + …

Related work

GCNZ



Related work

이 연구들은 전부 **매핑 함수**를 학습하는 다양한 방법론을 제시함.

그런데, 매핑의 대상이 되는 양쪽 벡터 공간의 도메인이 서로 매우 상이함(heterogeneous)

* 양쪽 벡터 공간 : Visual feature / 시맨틱 임베딩(워드 임베딩)

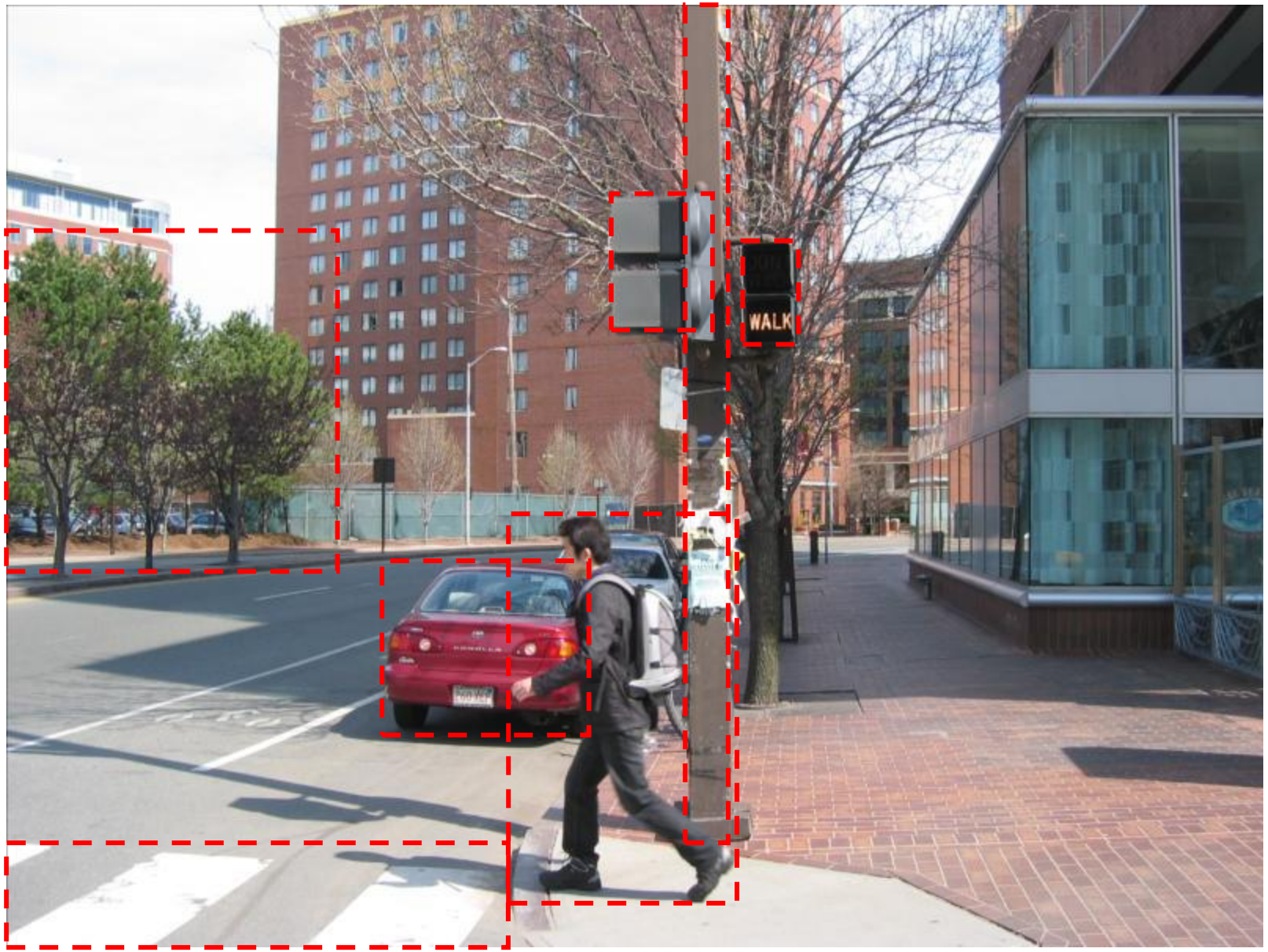
1) Visual feature

   low-level 이미지 데이터로부터 특정 태스크(classification 등)를 학습하며 자연스럽게 생성된 피쳐 추출기로부터 만들어진 벡터

2) 시맨틱 임베딩(워드 임베딩)

   대량의 **text 데이터**를 보며, 단어들이 함께 등장(text context)하는 횟수 등의 통계적 수치를 비지도학습을 통해 학습하여 만들어진 임베딩 벡터

Related work



Related work

서로 상이한 도메인 간의 매핑(**cross-domain mapping**)을 학습하는 것은 쉽지 않음

(domain shift problem 등)

Text context 로부터 학습한 워드 임베딩이 아니라,

Visual context 로부터 학습한 임베딩(Visual Context Embedding)을 사용하면,

매핑의 대상이 되는 양쪽 도메인이 homogeneous 하므로,

매핑 학습이 더 잘될 것이다!

Approach

Approach
Visual Context Embedding

Visual Context Embedding

Approach
Visual Context Embedding

워드 임베딩 방법론 중 GloVE 의 학습 방식과 유사

다만, 학습 데이터로 text corpora를 사용하는 것이 아니라
이미지 데이터셋의 annotation (label, bbox 등) 을 사용함

Approach
Visual Context Embedding

Annotation


```
{
    "image_id": 2,
    "objects": [
        {
            "center": [
                182.0,
                472.0
            ],
            "h": 254,
            "synsets": "road.n.01",
            "w": 364,
            "x": 0,
            "y": 345
        },
        {
            "center": [
                559.0,
                473.5
            ],
            "h": 253,
            "synsets": "sidewalk.n.01",
            "w": 478,
            "x": 320,
            "y": 347
        },
```


Co-occurrence matrix

X

	w_0	w_1	w_2	…
w_0				
w_1				
w_2				
…				

Approach
Visual Context Embedding

Co-occurrence matrix X

각 단어를 임의의 벡터로 초기화한 임베딩 스페이스 $w, \tilde{w}$

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 ,$$

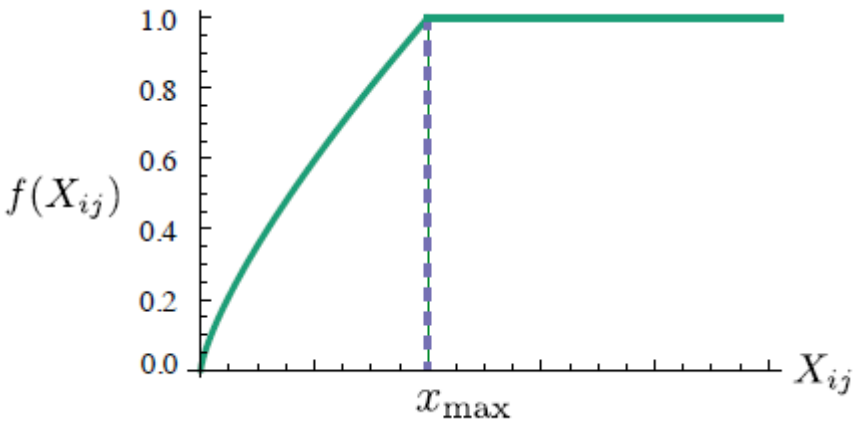
두 단어 간 dot product

(=코사인 유사도)

두 단어가 동시에 등장한

횟수의 log-scale 값

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} .$$



Approach
Visual Context Embedding

Detail

- 학습에 사용한 annotation 을 가져온 Dataset : Visual Genome
- 단어 사전 크기 : 3K (annotation에서 등장횟수 최소 10 이상)
- 벡터 사이즈 : 16차원/32차원/64차원/128차원

테스트 데이터셋

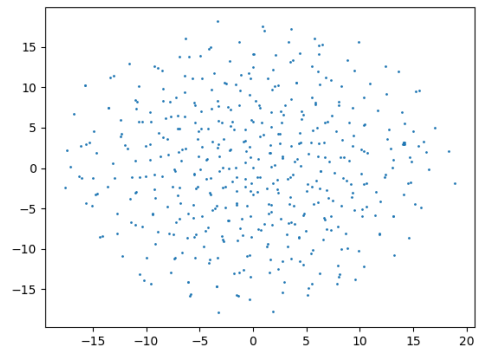
- Visual Genome (seen : unseen = 478 : 130)
- MS COCO (seen : unseen = 60 : 20)

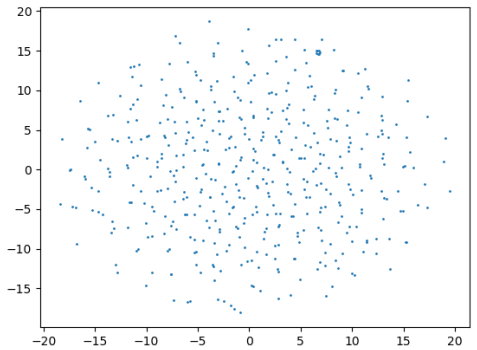
Experiment

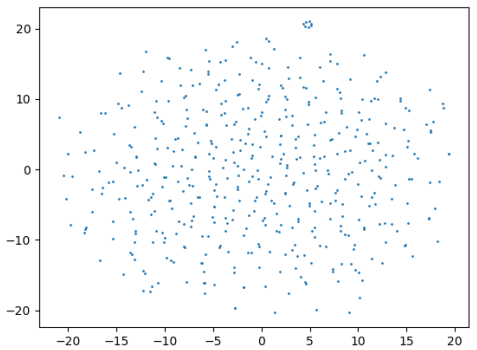
Experiment
Qualitative result

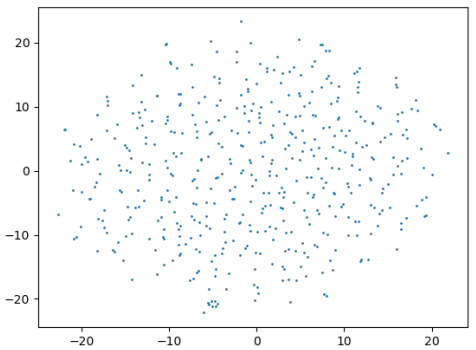
- Visual Context Embedding 시각화

bbox 미사용

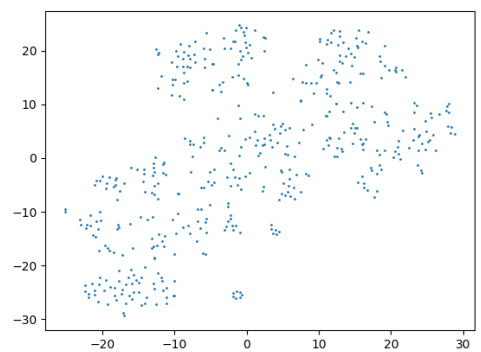


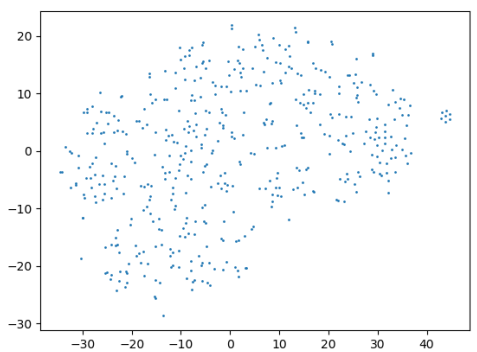


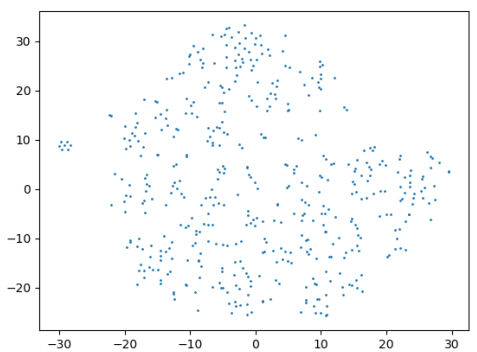


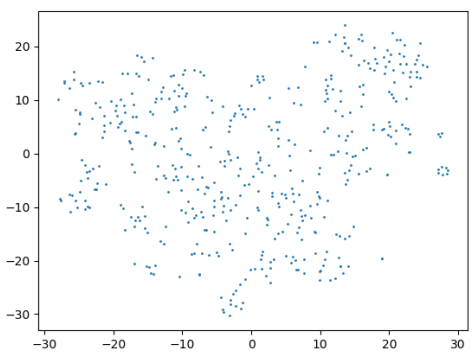


bbox 사용









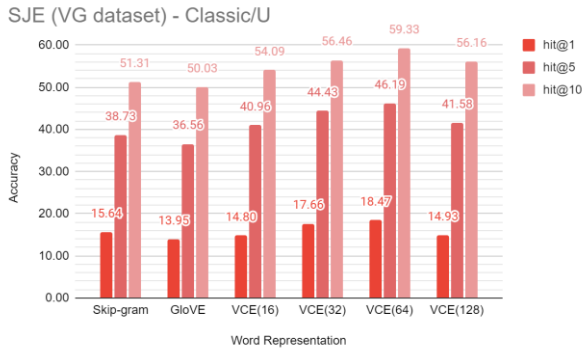
VCE(16차원) VCE(32차원) VCE(64차원) VCE(128차원)

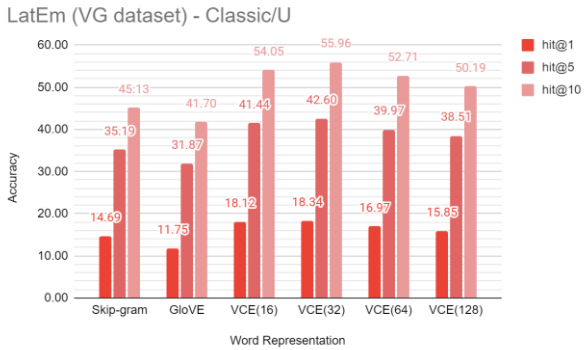
Experiment
Quantitative result

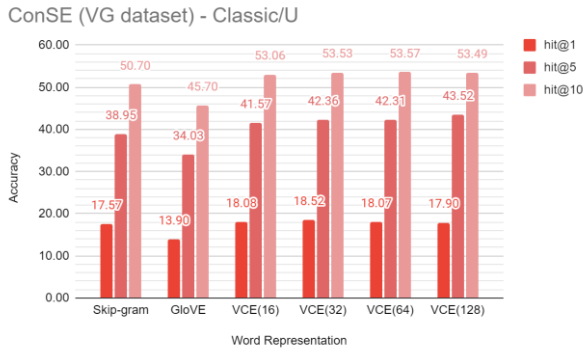
Visual Genome 데이터셋에 대한 성능 평가 결과

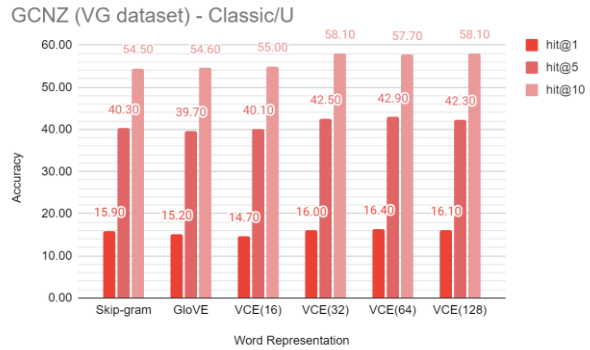
		Classic/U						Generalized/U					
		hit@1		hit@5		hit@10		hit@1		hit@5		hit@10	
		per-class	per-instance	per-class	per-instance	per-class	per-instance	per-class	per-instance	per-class	per-instance	per-class	per-instance
SJE	Skip-gram	15.64	20.38	38.73	**50.24**	51.31	61.35	2.78	5.37	14.33	23.72	24.69	**36.80**
	GloVE	13.95	**21.60**	36.56	47.46	50.03	59.38	4.06	**6.56**	16.52	**24.50**	24.95	33.55
	VCE(16)	14.80	14.30	40.96	44.82	54.09	61.12	4.09	3.46	17.59	16.25	27.05	28.57
	VCE(32)	17.66	17.89	44.43	50.18	56.46	**65.33**	**5.38**	4.29	**20.24**	20.22	**31.01**	34.06
	VCE(64)	**18.47**	16.97	**46.19**	45.96	**59.33**	60.69	4.60	4.33	18.90	18.06	28.72	30.17
	VCE(128)	14.93	12.92	41.58	39.38	56.16	55.79	3.32	3.15	16.10	12.57	25.69	22.53
LatEm	Skip-gram	14.69	**22.93**	35.19	44.74	45.13	56.37	2.88	**9.98**	13.38	20.90	21.53	29.83
	GloVE	11.75	15.84	31.87	39.82	41.70	50.19	2.38	4.12	10.78	16.03	19.51	26.39
	VCE(16)	18.12	21.63	41.44	53.12	**54.05**	65.62	4.20	5.03	18.55	23.94	**29.12**	37.52
	VCE(32)	**18.34**	21.73	**42.60**	**54.37**	55.96	**66.46**	**4.39**	5.03	**19.23**	**24.82**	28.94	**39.62**
	VCE(64)	16.97	19.22	39.97	46.82	52.71	59.44	3.39	4.20	16.96	17.47	26.28	30.43
	VCE(128)	15.85	11.71	38.51	37.22	50.19	50.49	2.35	2.09	12.11	10.21	20.79	20.50
ConSE	Skip-gram	17.57	23.33	38.95	46.78	50.70	58.14	0.17	0.26	14.43	20.07	23.70	30.86
	GloVE	13.90	28.21	34.03	53.18	45.70	64.33	0.21	1.50	12.98	28.20	21.67	40.20
	VCE(16)	18.08	30.31	41.57	62.75	53.06	74.74	**1.86**	**4.73**	18.48	33.96	28.74	47.91
	VCE(32)	**18.52**	**32.31**	42.36	**64.83**	53.53	**75.82**	1.30	4.23	**20.37**	**38.45**	**29.56**	**52.08**
	VCE(64)	18.07	24.97	42.31	62.80	**53.57**	75.39	0.35	1.06	17.81	25.11	28.57	46.23
	VCE(128)	17.90	18.18	**43.52**	57.36	53.49	72.01	0.06	0.14	15.57	14.40	27.18	31.66
ConSE(10)	Skip-gram	17.71	22.74	38.83	46.84	50.30	58.23	0.00	0.02	14.46	19.71	24.32	31.39
	GloVE	14.23	27.14	34.82	52.27	45.91	63.67	0.04	0.25	12.67	26.30	22.00	38.28
	VCE(16)	18.46	28.96	42.36	60.70	53.47	72.45	**1.30**	**2.76**	19.03	31.25	29.63	45.49
	VCE(32)	**19.36**	**31.43**	42.72	**62.82**	53.95	**73.75**	0.83	2.11	**20.78**	**35.98**	**30.29**	**49.37**
	VCE(64)	18.61	24.00	42.68	60.64	**54.20**	72.93	0.15	0.37	18.04	22.82	29.18	43.29
	VCE(128)	18.20	17.38	**43.34**	54.44	53.63	69.15	0.02	0.03	15.92	13.43	27.76	29.94
ConSE(100)	Skip-gram	17.71	23.36	39.18	47.15	50.90	58.79	0.13	0.20	14.65	20.14	24.02	31.16
	GloVE	14.04	28.23	34.26	53.23	45.94	64.55	0.18	1.32	13.06	28.15	21.88	40.09
	VCE(16)	18.28	30.29	41.91	62.72	53.59	74.56	**1.77**	**4.48**	18.56	33.70	28.93	47.72
	VCE(32)	**18.68**	**32.40**	42.81	**64.72**	53.81	**75.67**	1.24	3.99	**20.54**	**38.37**	**29.73**	**51.74**
	VCE(64)	18.18	24.95	42.58	62.84	**54.04**	75.26	0.31	0.93	17.85	24.87	28.66	45.96
	VCE(128)	18.01	18.09	**43.85**	57.14	53.81	71.91	0.05	0.10	15.69	14.25	27.25	31.43
GCNZ	Skip-gram	15.90	23.50	40.30	49.50	54.50	62.90	7.03	12.60	15.00	24.40	24.60	33.70
	GloVE	15.20	20.40	39.70	46.40	54.60	61.50	7.09	11.00	15.30	23.30	24.40	31.70
	VCE(16)	14.70	22.10	40.10	51.00	55.00	65.10	6.59	10.40	16.30	25.10	24.60	35.30
	VCE(32)	16.00	**24.90**	42.50	**53.00**	**58.10**	**68.40**	7.59	**14.00**	17.60	**27.90**	27.00	**38.00**
	VCE(64)	**16.40**	22.80	**42.90**	52.00	57.70	67.60	**7.80**	12.90	**18.40**	26.90	**28.00**	37.30
	VCE(128)	16.10	22.50	42.30	51.50	58.10	66.20	7.57	10.40	16.50	25.40	26.20	35.80

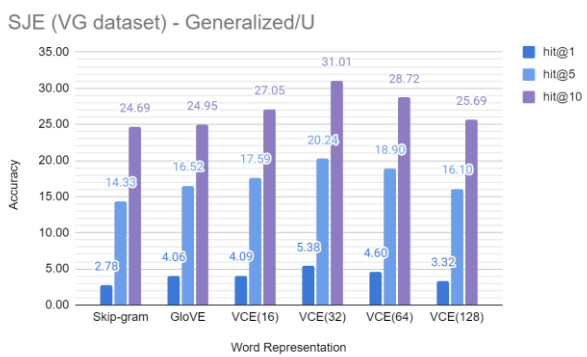
Experiment
Quantitative result

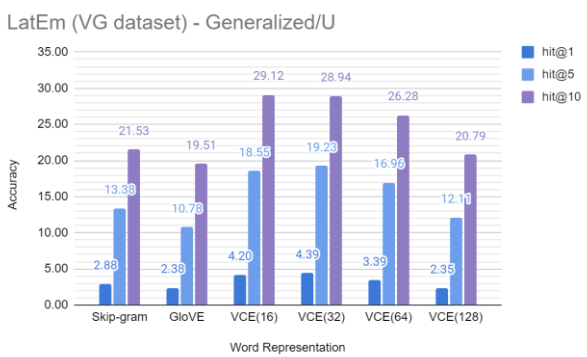


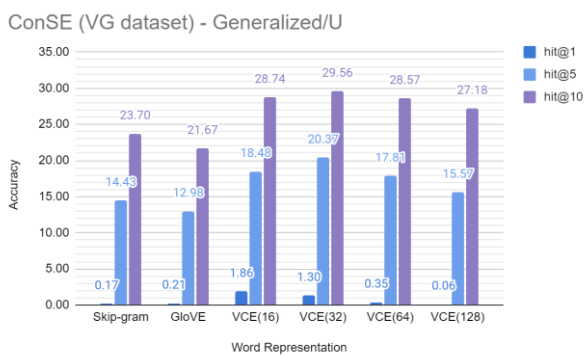


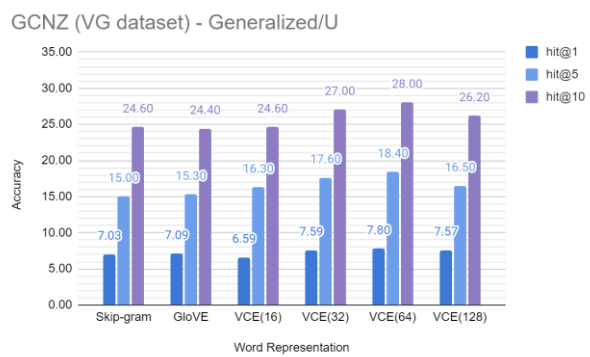












Visual Genome 데이터셋에 대한 성능 평가 결과

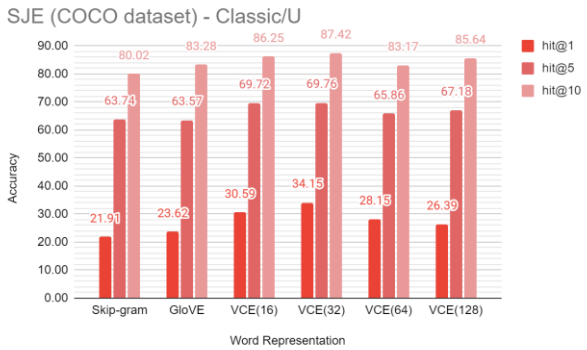
(per-class accuracy)

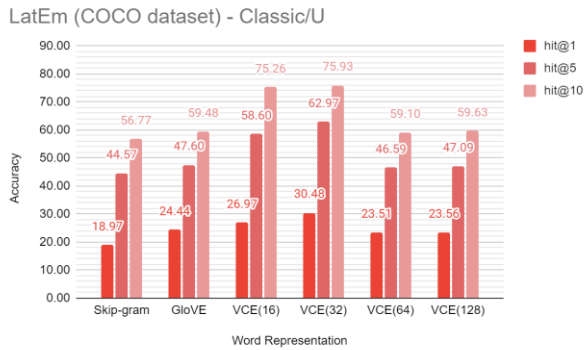
Experiment
Quantitative result

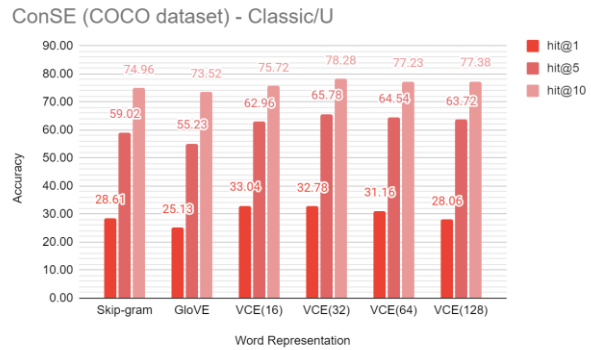
MS COCO 데이터셋에 대한 성능 평가 결과

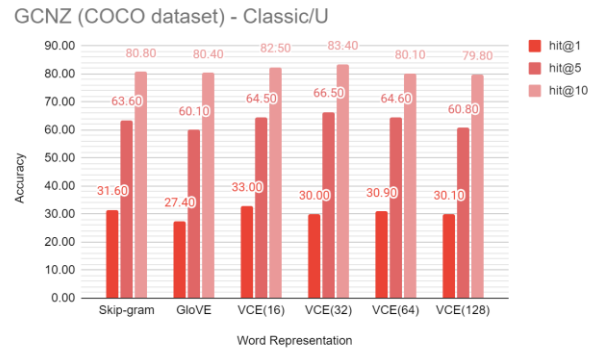
		Classic/U						Generalized/U					
		hit@1		hit@5		hit@10		hit@1		hit@5		hit@10	
		per-class	per-instance	per-class	per-instance	per-class	per-instance	per-class	per-instance	per-class	per-instance	per-class	per-instance
SJE	Skip-gram	21.91	20.33	63.74	59.63	80.02	77.56	2.85	4.09	23.67	26.01	45.89	44.03
	GloVE	23.62	20.49	63.57	56.81	83.28	77.82	3.82	3.15	24.33	21.37	42.84	37.74
	VCE(16)	30.59	27.92	69.72	**65.34**	86.25	**83.98**	8.98	7.36	36.47	33.63	53.75	49.19
	VCE(32)	**34.15**	**28.83**	**69.76**	63.59	**87.42**	80.35	**12.03**	**10.86**	**40.71**	**36.15**	**55.35**	**50.19**
	VCE(64)	28.15	23.57	65.86	58.95	83.17	79.41	6.52	6.42	34.39	31.39	52.45	47.28
	VCE(128)	26.39	21.69	67.18	60.51	85.64	81.74	3.16	2.37	27.98	22.50	45.23	38.42
LatEm	Skip-gram	18.97	18.71	44.57	39.20	56.77	50.91	1.47	2.08	14.08	14.72	28.64	24.90
	GloVE	24.44	21.24	47.60	42.41	59.48	54.28	2.11	1.88	16.59	14.49	29.91	26.33
	VCE(16)	26.97	25.94	58.60	55.84	75.26	**71.50**	**4.04**	**4.57**	30.35	**29.57**	**45.76**	**43.03**
	VCE(32)	**30.48**	**27.33**	**62.97**	**57.33**	**75.93**	69.62	3.44	3.99	**30.55**	28.73	45.76	41.93
	VCE(64)	23.51	20.95	46.59	44.36	59.10	56.39	2.56	2.43	20.27	18.22	34.26	30.45
	VCE(128)	23.56	17.80	47.09	38.49	59.63	50.32	0.31	0.29	9.14	8.07	23.83	19.03
ConSE	Skip-gram	28.61	24.81	59.02	50.84	74.96	62.42	0.67	**1.17**	25.75	22.80	42.30	37.26
	GloVE	25.13	22.89	55.23	48.05	73.52	68.45	0.00	0.00	20.86	19.16	33.43	29.80
	VCE(16)	**33.04**	28.79	62.96	58.40	75.72	**73.18**	**0.73**	0.62	33.88	30.84	48.75	45.53
	VCE(32)	32.78	**29.18**	**65.78**	**58.85**	**78.28**	69.88	0.15	0.13	**36.43**	**33.43**	**50.55**	**46.43**
	VCE(64)	31.16	27.72	64.54	58.75	77.23	69.13	0.03	0.03	33.29	30.38	48.84	44.94
	VCE(128)	28.06	24.45	63.72	58.43	77.38	71.69	0.00	0.00	26.75	24.94	45.10	40.73
ConSE(10)	Skip-gram	28.94	25.03	59.12	50.62	75.10	62.48	0.67	**1.17**	26.12	23.05	43.08	37.68
	GloVE	25.59	23.12	55.93	48.87	74.08	68.94	0.00	0.00	21.17	19.42	34.17	30.29
	VCE(16)	33.32	28.99	63.22	58.79	75.98	**73.25**	**0.81**	0.68	33.99	30.97	48.70	45.27
	VCE(32)	**33.53**	**29.18**	**66.24**	**59.11**	**78.39**	69.84	0.21	0.16	**36.41**	**33.46**	**51.19**	**46.95**
	VCE(64)	31.28	27.79	64.67	58.63	78.09	69.62	0.03	0.03	33.30	30.45	48.85	44.78
	VCE(128)	28.12	24.58	63.99	58.59	77.79	72.11	0.00	0.00	26.97	25.16	45.22	40.86
ConSE(100)	Skip-gram	31.60	27.00	63.60	53.40	80.80	71.50	5.69	5.38	31.40	26.40	46.20	38.90
	GloVE	27.40	23.80	60.10	52.60	80.40	74.00	5.51	5.21	28.90	24.40	43.70	36.60
	VCE(16)	**33.00**	**28.30**	64.50	55.50	82.50	77.50	7.11	6.16	**33.60**	**28.30**	48.50	41.20
	VCE(32)	30.00	25.30	**66.50**	**56.50**	**83.40**	**77.70**	**9.23**	**8.11**	33.50	28.20	**49.00**	**42.70**
	VCE(64)	30.90	26.10	64.60	54.30	80.10	72.70	5.66	4.99	30.00	25.60	48.70	40.40
	VCE(128)	30.10	24.90	60.80	54.00	79.80	75.60	4.37	4.11	24.90	21.60	42.40	36.30

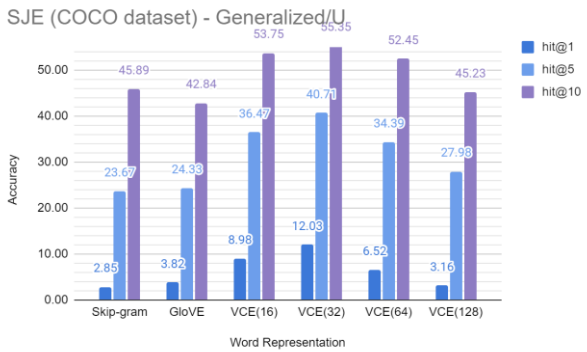
Experiment
Quantitative result

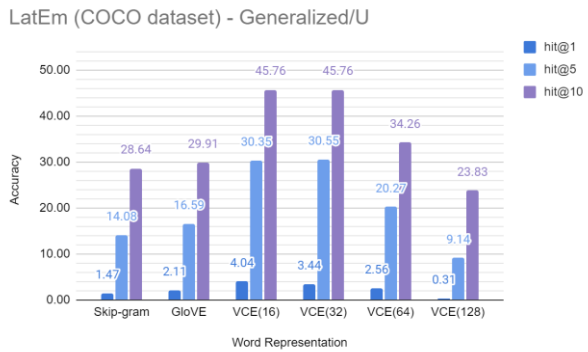


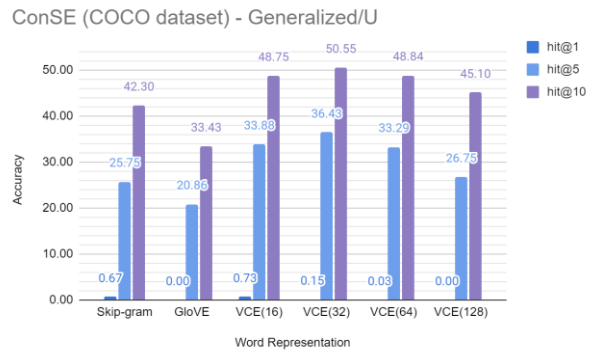


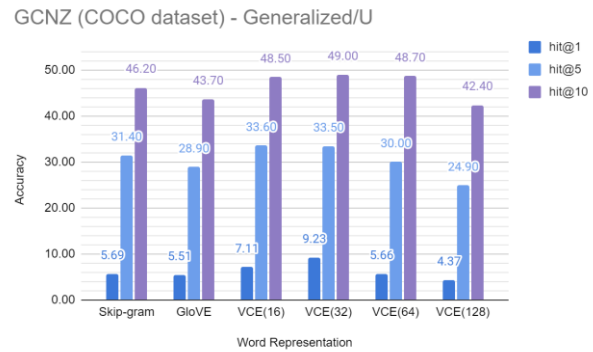












MS COCO 데이터셋에 대한 성능 평가 결과

(per-class accuracy)

Conclusion

Conclusion
Contribution

1) 이 연구에서는 Zero-Shot Recognition 태스크에 **새로운 임베딩 스페이스(Visual Context Embedding)**를 적용하여 기존 기법의 성능을 향상시킴.
2) 기존 ZSR 패러다임의 핵심인 상이한 도메인 간 매핑에 근본적인 문제가 있음을 밝히고, 매핑 함수의 학습 성능 개선이 아닌, 매핑의 도메인 자체를 바꾸기 위한 새로운 시도를 함.

Q & A

Thank you