

# Visual Word2Vec(vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes

2016, CVPR, S. Kottur et al.

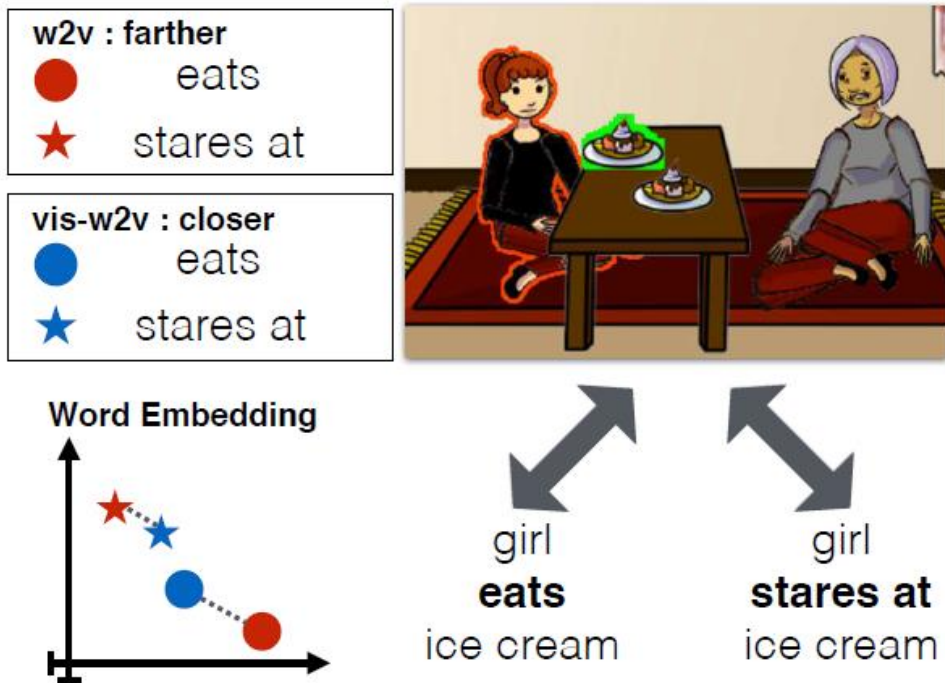
한양대학교  
컴퓨터 소프트웨어 학과  
인공지능 연구실  
조건희

# Introduction

---

# Introduction

- 이 논문에서 태클하는 문제



- “eat” 과 “stare at” 은 텍스트 상으로만 보면 관련이 없을 것 같지만,
- 이미지에서는 서로 연관이 있을 가능성이 있다.

- 이 논문에서 태클하는 문제

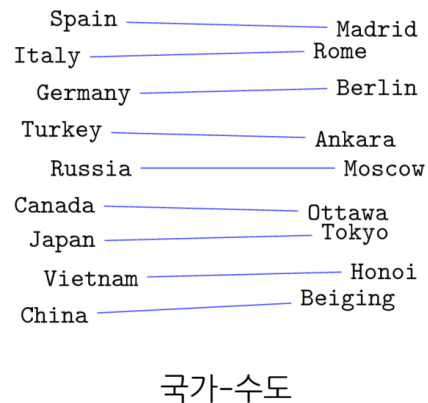
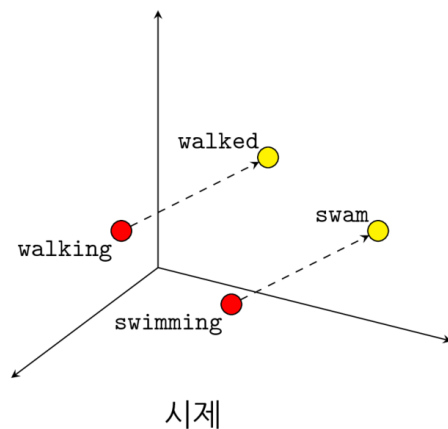
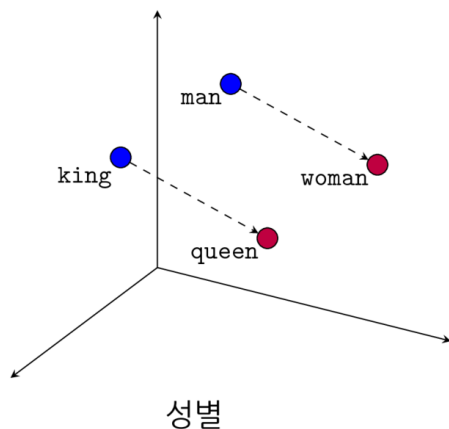
- 그래서 클립아트 이미지 데이터를 활용하여 **visual grounding** 을 해보겠다!
- (= 텍스트만으로는 학습할 수 없지만, 이미지에서는 학습할 수 있을 법한 단어 간 연관성을 찾아 보겠다!)
- 그리고 워드 임베딩이 잘 되었는지 확인하기 위해 3가지 태스크를 준비했다!
  - Common sense assertion classification
    - (boy, eats, cake) 형태로 주어지는 문장이 타당한지(상식적인지) 판별하는 태스크 → SOTA를 찍었다!
  - Visual paraphrasing
    - 주어진 2개의 문장이 같은 장면을 묘사하는지 아닌지(서로에 대한 paraphrase 인지 아닌지) 판별하는 태스크
  - Text-based image retrieval
    - 문장이 주어지면 그에 해당하는 이미지를 찾는 태스크

# Related work

## Related work

### ▪ Word embedding

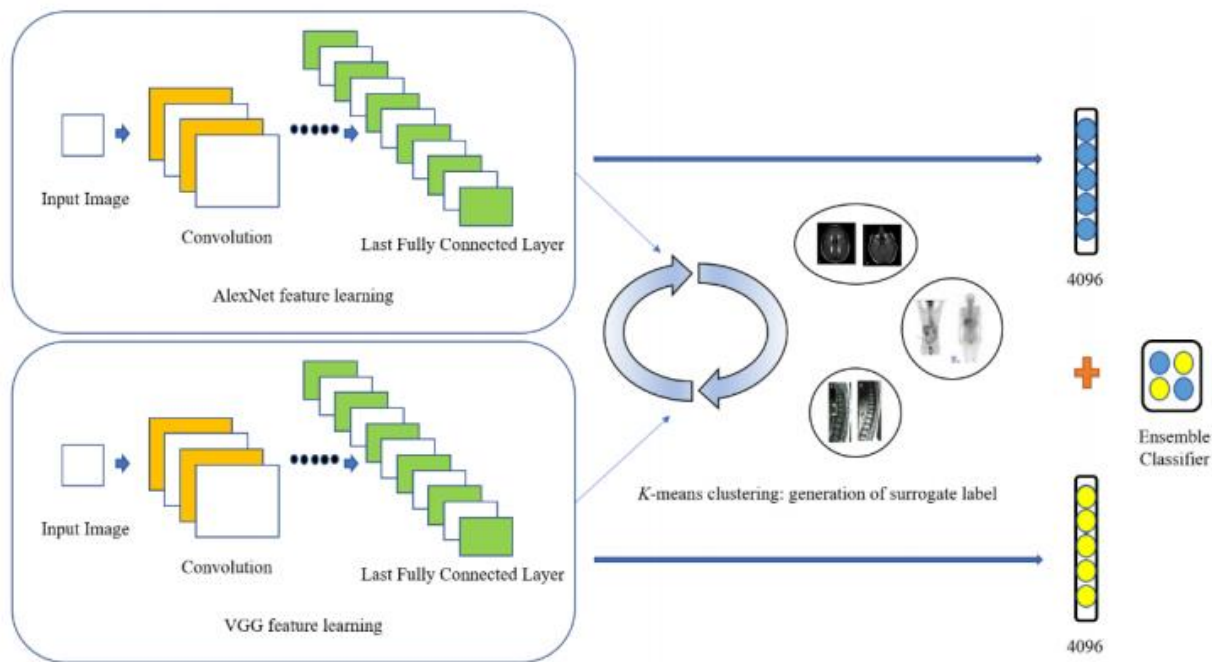
- 대량의 텍스트 데이터에서 단어들의 co-occurrence 를 학습하여 서로 연관있는 단어들끼리는 가깝도록, 연관없는 단어들끼리는 멀도록 임베딩하는 것
- 임베딩된 단어 벡터들간의 유사도(코사인유사도 등)은 그 단어들 간의 의미적 유사도로 사용가능!



## Related work

### ▪ Surrogate classification

- unsupervised learning 에서 주로 사용.
- 인풋 이미지에 대한 라벨이 없을 때, 일단 그 이미지의 feature 를 뽑아 라벨이 있는 feature와의 clustering을 통해 surrogate label (대리 라벨) 을 할당해줌.



# Approach



# Approach

## ▪ Input

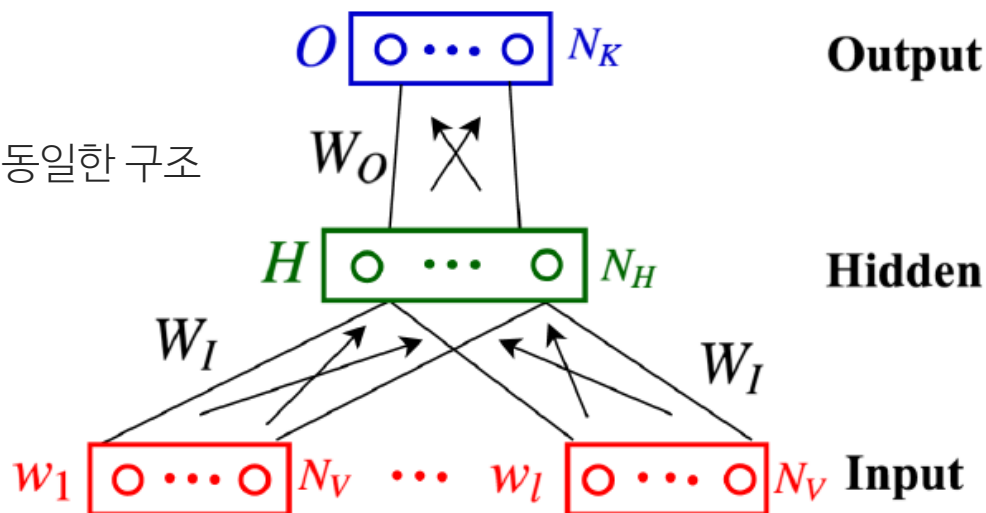
- 이미지 visual feature
- 이미지와 연관된 word set

## ▪ Model

- word2vec 기본 모델 중 CBOW 와 동일한 구조
- output layer 만 다름

## ▪ Output Classes

- Grounding function  $G(\cdot)$
- $G: v \rightarrow \{1, 2, \dots, N_K\}$
- 이미지에 포함된 단어를  $N_K$  개의 클러스터로 분류한 결과를 사용 (원핫벡터)

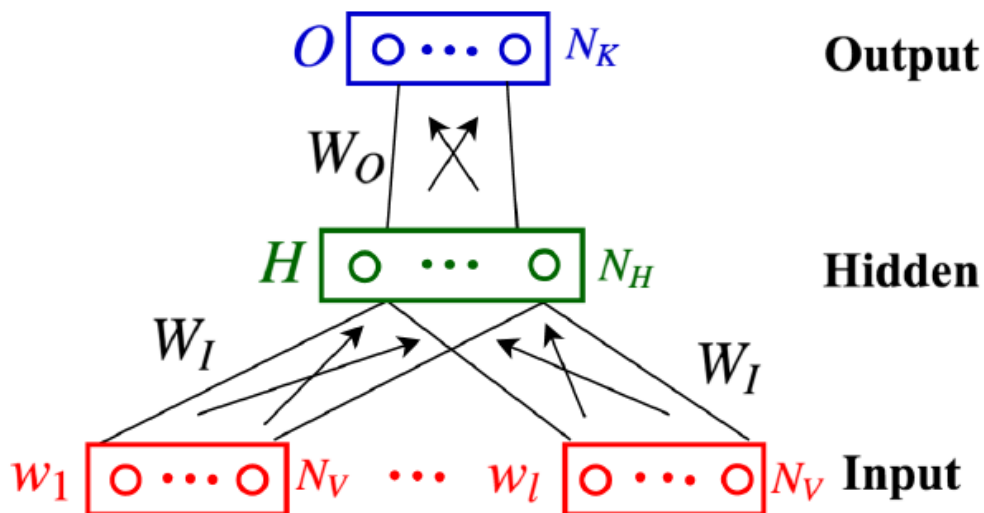




# Approach

- Initialization

- $W_I$ : 기존 w2v의 임베딩을 초기값으로 사용
- $W_O$ : 랜덤값
- 초기값 자체가 이미 대량의 텍스트로부터 학습한 임베딩이기 때문에 non-visual 한 정보 반영



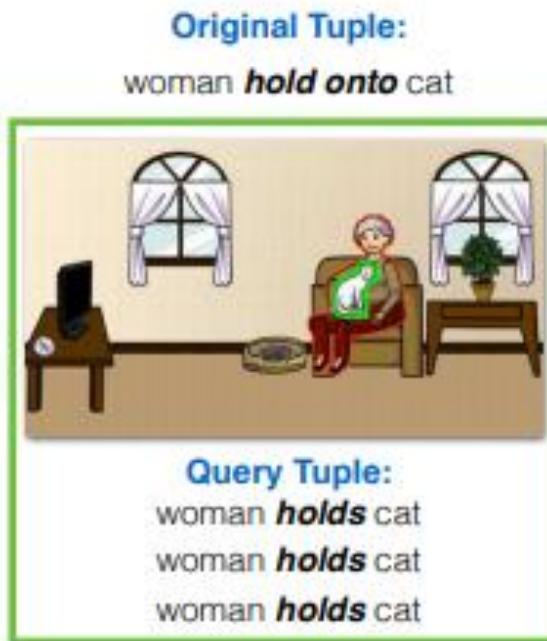
# Experiments

# Experiments

## ▪ Common Sense Assertion Classification 태스크

### ▪ 데이터셋 수집

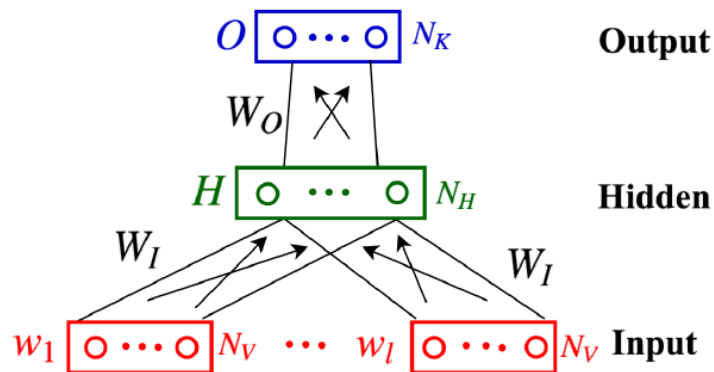
- 이미지를 묘사하는 문장 데이터셋을 AMT (Amazon Mechanical Turk) 이용하여 수집
- 각 문장을 (주어, 동사, 목적어) 형태



# Experiments

## ▪ Common Sense Assertion Classification 태스크

- 이미지를 묘사하는 문장이 타당한지 아닌지 구분하는 태스크
- 이미지에 대한 Ground truth tuple이 존재함.
- 입력으로 tuple이 하나 주어짐.  $(t_p, t_r, t_s)$  (예를 들면, (boy, eats, cake))
- 테스트 방법: plausible 하다고 알려져 있는 tuple과의 유사도를 워드임베딩을 사용해 측정
- plausibility score 계산 :  $h(t', t_i) = W_P(t'_P)^T W_P(t_P^i) + W_R(t'_R)^T W_R(t_R^i) + W_S(t'_S)^T W_S(t_S^i)$
- separate model / shared model
  - shared model은  $W_P = W_R = W_S$



# Experiments


- Common Sense Assertion Classification 태스크
  - w2v-wiki : 위키피디아 텍스트로부터 학습한 워드임베딩
  - w2v-coco : MS-COCO 데이터셋의 이미지 caption 으로부터 학습한 워드임베딩
  - vis-w2v-wiki : 초기값이 wiki 워드임베딩
  - vis-w2v-coco : 초기값이 MS-COCO 워드임베딩
  - AP : Average precision

Approach	common sense AP (%)
vis-w2v-wiki (shared)	72.2
vis-w2v-wiki (separate)	74.2
vis-w2v-coco (shared) + vision	74.2
vis-w2v-coco (shared)	74.5
vis-w2v-coco (separate)	<b>74.8</b>
vis-w2v-coco (separate) + vision	<b>75.2</b>
w2v-wiki (from [35])	68.4
w2v-coco (from [35])	72.2
w2v-coco + vision (from [35])	73.6

# Experiments

## ▪ Visual paraphrasing 태스크

- 이미지를 묘사하는 두 텍스트가 같은 장면을 묘사하고 있는지 아닌지 판별하는 태스크
- 텍스트에 포함된 모든 단어 벡터의 평균으로 text-based scoring function에 넣어 paraphrasing score를 판별
- text-based scoring function : term freq, word co-occurrence 등을 결합한 함수

Jenny is kicking Mike. Mike dropped the soccer ball on the duck. There is a sandbox nearby.		Mike and Jenny are surprised. Mike and Jenny are playing soccer. The duck is beside the soccer ball.
Mike is in the sandbox. Jenny is waving at Mike. It is a sunny day at the park.		Jenny is very happy. Mike is sitting in the sand box. Jenny has on the color pink.
Mike and Jenny say hello to the dog. Mike's dog followed him to the park. Mike and Jenny are camping in the park.		The cat is next to Mike. The dog is looking at the cat. Jenny is waving at the dog.



# Experiments

---

- Visual paraphrasing 태스크

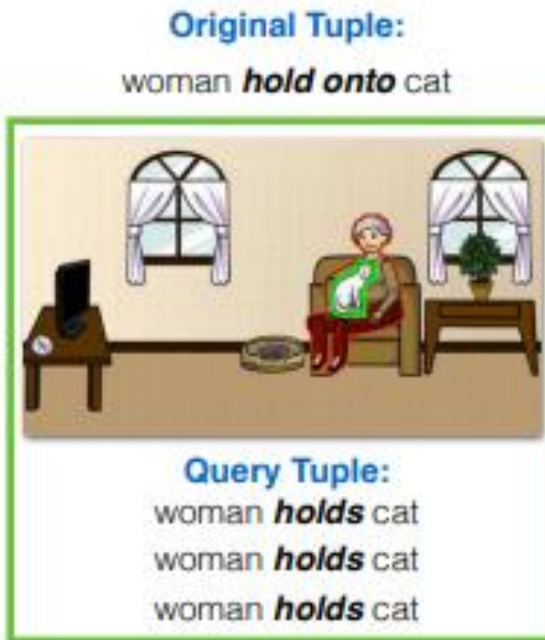
- 이미지를 묘사하는 두 텍스트가 같은 장면을 묘사하고 있는지 아닌지 판별하는 태스크
- 텍스트에 포함된 모든 단어 벡터의 평균으로 text-based scoring function에 넣어 paraphrasing score를 판별
- text-based scoring function : term freq, word co-occurrence 등을 결합한 함수

Approach	Visual Paraphrasing AP (%)
w2v-wiki (from [24])	94.1
w2v-wiki	94.4
w2v-coco	94.6
vis-w2v-wiki	95.1
vis-w2v-coco	95.3

# Experiments

## ▪ Text-based Image Retrieval 태스크

- query tuple 이 주어지면 그 query tuple 과 일치하는 ground truth tuple을 찾는 태스크
- 사실상 이미지랑은 크게 연관이 없는 태스크로 보임



# Experiments

## ▪ Text-based Image Retrieval 태스크

- query tuple 의 임베딩 벡터의 평균값과 ground truth tuple 의 평균과의 코사인 유사도로 판별.
- R@1 : Recall@1 (높을수록 좋음)
- med R : Median Rank (낮을수록 좋음)

Approach	R@1 (%)	R@5 (%)	R@10 (%)	med R
w2v-wiki	14.6	34.4	45.4	13
w2v-coco	15.3	35.2	47.6	11
vis-w2v-wiki (shared)	15.5	37.2	49.3	<b>10</b>
vis-w2v-coco (shared)	<b>15.7</b>	<b>37.7</b>	47.6	<b>10</b>
vis-w2v-wiki (separate)	14.0	32.7	43.5	15
vis-w2v-coco (separate)	15.4	37.6	<b>49.5</b>	<b>10</b>

# Conclusion

---

- Contribution

- 워드임베딩을 이미지 관련 태스크에 적용할 때 생길 수 있는 문제에 대한 문제제시
- visual information 을 워드임베딩과 콜라보하려 시도
- 3가지 태스크에서 기존 워드 임베딩보다 좋은 성능

감사합니다

---