공학석사학위논문

시각적 유사도와 의미적 유사도 간 비율 제어가 가능한 이미지 검색

Image-to-Image retrieval with controlling between visual and semantic similarity

2023 년 2 월

서울대학교 대학원

산업공학과

이 성 은

시각적 유사도와 의미적 유사도 간 비율 제어가 가능한 이미지 검색

Image-to-Image retrieval with controlling between visual and semantic similarity

지도교수 이 경 식

이 논문을 공학석사 학위논문으로 제출함

2023 년 2월

서울대학교 대학원

산업공학과

이 성 은

이성은의 공학석사 학위논문을 인준함

2023 년 2 월

위 원 장 조 성 준 (인)

부위원장 이 경 식 (인)

위 원 박 종 헌 (인)

초록

이미지 검색(Image-to-Image retrieval)은 쿼리(query) 이미지에 대해 유사한 이미지를 찾아주는 작업으로, 주로 크게 시각적 유사도와 의미적 유사도의 두 가지 접근 방식으로 나뉘어 연구되어왔다. 이미지 검색은 다양한 상황에서 이루어지기 때문에 하나의 기준으로 검색하는 것은 사용자의 의도에 유연하게 대응하기에는 한계가 있다.

본 논문에서는 시각적 관점과 의미적 관점 모두를 고려하여, 사용자가 목적에 따라 자유롭게 비중을 조절하고 사용자의 의도에 따라 유연하게 검색할 수 있는 방법론을 제안한다. 이를 위해 이미지를 표현하는 장면 그래프(Scene graph)에서 그래프 합성곱 신경망(Graph Convolutional Network)을 통해 시각적 특징과 의미적 특징을 추출한 후, 이를 내삽(Interpolation)하여 비율을 조절하여 검색하는 모델이다. 이미지에서 사전 학습된(pre-trained) ResNet-152 모델을 통해 추출한 시각적 특징, 사람이 작성한 캡션에서 사전 학습된 Sentence-Bert(SBERT) 모델을 통해 추출한 의미적 특징을 대리 관련도(Surrogate relevance)로 활용하여 학습을 했다.

이를 통해 학습한 시각적·의미적 특징들이 각각 학습한 대리 관련도 측면에서 높은 normalized Discounted Cumulative Gain(nDCG)를 보임으로써 그래프 레벨을 통해 각 특징을 성공적으로 추출했음을 보였다. 또한, 알고리즘이 사람의 평가와 얼마나 유사한지를 보여주는 인간 동의 점수(Human agreement score)에서 다른 선행 연구들과 정량적 성능 비교를 통해 이미지 검색 성능이 뛰어나다는 것을 검증할 수 있었다. 또한, 시각적·의미적 특징을 ResNet 및 캡션 SBERT로 대체한 모델과 비교 결과, 같은 그래프 레벨 상에서 추출한 특징을 이용한 내삽이 더 좋은 성능을 보임을 확인할 수 있었다. 이미지에 대해 비율을 조정하며 검색한 정성적 결과를 통해 본 모델이 성공적으로 시

각적·의미적 유사도의 비율을 조정한 검색을 수행할 수 있음 역시 확인하였다.

주요어: 이미지 검색, 이미지 간 유사도, 이미지 시각적 유사도, 이미지 의미적 유사도, 장면 그래프, 그래프 임베딩, 그래프 합성곱 신경망, 대리 관련도, 인간 동의 점수, ResNet, Sentence BERT

학번: 2019-27796

목차

초록 **i**

목차 **v**

표 목차 **vi**

그림 목차 **vii**

제 1 장 서론 **1**

1.1 연구의 배경 및 동기 1

1.2 연구 목적 3

1.3 연구 공헌 4

1.4 논문구성 5

제 2 장 배경 이론 및 관련 연구 **6**

2.1 배경 이론 6

2.1.1 장면 그래프 6

2.1.2 그래프 합성곱 신경망 9

2.2 관련 연구 14

2.2.1 이미지 특징 추출 연구 14

2.2.2 이미지 검색 15

제 3 장 제안 기법 **17**

3.1 이미지 대리 관련도 추출 18

3.1.1 시각적 대리 관련도 18

3.1.2 의미적 대리 관련도 19

3.2 장면 그래프를 통한 이미지 간 유사도 계산 20

3.2.1 시각적 특징 추출 그래프 21

3.2.2 의미적 특징 추출 그래프 21

3.2.3 모델 학습 22

3.2.4 시각적·의미적 유사도의 비율을 반영한 이미지 간 유사도 추론 . 23

제 4 장 실험 결과 **24**

4.1 데이터셋 24

4.1.1 VG-COCO 24

4.1.2 캡션 및 장면 그래프 생성 27

4.1.3 인간 동의 점수 30

4.2 실험 세팅 32

4.2.1 2단계 이미지 검색 32

4.2.2 학습 세팅 32

4.3 실험 결과 33

4.3.1 정량적 실험 결과 33

4.3.2 정성적 실험 결과 42

제 5 장 결론 **44**

5.1 결론 44

5.2 향후 연구 45

참고문헌 **46**

Abstract **55**

표 목차

표 4.1 정답 장면 그래프를 이용해 이미지 검색을 진행한 결과로 ResNet-152(시각적 대리 관련도에 사용한 모델) 유사도를 정답 라벨로 간주했을 때의 ndCG 결과 35
표 4.2 정답 장면 그래프를 이용해 이미지 검색을 진행한 결과로 정답 캡션 SBERT(의미적 대리 관련도에 사용한 모델)을 정답 라벨로 간주했을 때의 ndCG 결과 36
표 4.3 정답 장면 그래프를 사용했을 때의 인간 동의 점수 37
표 4.4 생성된 장면 그래프를 사용했을 때의 인간 동의 점수 38

그림 목차

그림 2.1 이미지와 대응되는 경계 상자 및 장면 그래프 예시 8
그림 2.2 그래프 합성곱 신경망 구조 11

그림 3.1 본 연구에서 제안하는 VvsS-Net 학습 방식 20

그림 4.1 학습 데이터 이미지 예시 25
그림 4.2 4.1의 Visual Genome 라벨 26
그림 4.3 4.1의 MS-COCO segmentation 라벨 27
그림 4.4 테스트 데이터 이미지 예시 28
그림 4.5 4.4의 생성된 경계 상자 및 장면 그래프 29
그림 4.6 이미지 트리플렛 예시 30
그림 4.7 시각적 특징과 의미적 특징의 내삽 비율에 따른 인간 동의 점수 . 39
그림 4.8 시각적 유사도 비중에 따른 높은 인간 동의 점수를 보인 사람 수 41
그림 4.9 다른 λ 값에 따른 이미지 검색 결과 예시 43

제 1 장 서론

1.1 연구의 배경 및 동기

이미지 검색 작업(Image-to-Image retrieval)은 컴퓨터 비전 분야에서 가장 중요한 연구 주제 중 하나로, 주어진 쿼리(query) 이미지에서 잠재적으로 큰 이미지 데이터베이스 내에서 해당 쿼리와 관련된 비슷한 이미지를 검색하는 것을 목표로 한다. 검색 엔진들에서 필수적으로 갖추고 있는 기능으로, 기하급수적으로 늘어나는 이미지들과 멀티미디어 콘텐츠들 사이에서 사용자가 원하는 의도에 맞는 이미지를 찾아주는 요구는 계속되고 있다. 비슷한 랜드마크, 옷, 가구, 사람 식별과 같은 객체 위주의 검색(Instance-level search) 뿐만 아니라 이미지 속 객체 간의 관계까지 고려를 해야 하는 검색 등 이미지 검색의 양상이 다양해지고 복잡해지고 있다.

시각적 특징에 집중한 이미지 검색은 딥러닝 기술의 발달 이전에도 계속 활발히 연구되어왔던 유구한 분야로, Content Based Image Retrieval(CBIR)이라고도 불린다. 주로 색상, 질감, 모양 등 시각적 특징에 집중한 결과를 반환하며, 크게 객체 위주의 검색과 카테고리 위주의 검색을 목표로 한다 [1]. Fisher 벡터나 VLAD 설명자 [2]와 같은 특징 인코딩을 사용한 연구들이 전통적으로 진행되어왔고, 최근 딥러닝 기술의 발달로 합성곱 신경망(Convolutional Neural Network) 모델이 대두하며 이미지 검색 분야에서 효과성과 효율성 모두 대폭 향상되었다 [3, 4, 5]. 그러나 Convolutional Neural Network(CNN)을 이용한 이미지 검색의 경우 저수준 특징과 지역적 특징에 민감하게 반응하는 특징이 있다 [6, 7, 8]. 그렇기에 이미지 전체의 복잡한 관계를 고려한 검색 등에는 취약한 단점이 있다.

이미지 장면의 의미적 특징에 집중한 이미지 검색은 주로 이미지를 설명하는 텍스트 라벨인 캡션(Caption) [9], 이미지를 개체와 개체 간 관계를 통해 그래프로 표현한 장면 그래프(Scene graph) [10, 11, 12, 13]를 이용한 방법들이 연구가 되어왔다. 그러나 이러한 방법론들은 캡션이나 장면 그래프의 개체를 단어 라벨로 표현한 텍스트에 의존적이기에, 이미지의 시각적 특징을 잘 반영하지 못한다는 한계점이 있다.

이미지 검색에서는 시각적 특징에 집중한 연구, 이미지를 표현하는 장면의 의미적 특징에 집중한 연구는 모두 활발히 진행되고 있으나 이 둘 모두를 활용해서 상황에 맞는 방법론을 제안하는 연구는 부족한 상황이다. 본 연구에서는 기존에 발표된 장면 그래프를 이용한 의미적 특징에 집중한 검색 연구 [10]에서 제안한 방식을 확장시켜 시각적·의미적 특징 모두를 활용한 검색 프레임워크를 제안한다.

1.2 연구 목적

사용자의 의도에 따라, 혹은 이미지의 특성에 따라 다양한 상황과 맥락에서 이미지 검색이 이루어지기 때문에 시각적·의미적 유사도의 단일 기준으로 검색하는 것은 한계가 있다. 그렇기에 본 연구에서는 장면 그래프를 활용해 이미지로부터 동일한 수준에서의 시각적 특징과 의미적 특징을 추출하는 것을 목표로 한다. 그리고 이를 통해 이들 간의 비율을 내삽(Interpolation)을 통해 사용자의 의도에 맞게 시각적·의미적 유사도의 비율을 제어 가능하게 검색하는 방법론을 제안한다.

MS-COCO [14]와 Visual-Genome [15] 데이터셋을 통해 캡션과 장면 그래프가 모두 있는 데이터셋을 구축하고 해당 데이터셋에서 이미지 간 유사도를 라벨링한 인간 동의 점수(Human agreement score) 데이터를 이용했다. 대리 관련도를 통해 장면 그래프의 시각적·의미적 특징을 학습하여 추출하고 이들이 각각 이미지의 시각적·의미적 특징을 잘 표현할 수 있도록 학습했는지를 확인한다. 또한, 인간 동의 점수 측면에서 두 가지 측면을 모두 고려하는 것이 검색 성능을 높일 수 있음을 확인한다. 절제 연구를 통해 장면 그래프의 같은 수준에서의 특징 추출이 제어 및 검색 성능에 긍정적 효과가 있음을 확인한다. 이를 통해 사용자의 의도 또는 이미지의 특성에 맞게 이미지의 시각적·의미적 비율을 제어한 검색 방법론을 제안한다.

1.3 연구 공헌

본 연구의 동기 및 공헌은 다음과 같다.

(a) 기존 모델에 제어 가능한 단위 모듈을 추가하여 시각적·의미적 유사도 간의 비율을 제어할 수 있는 이미지 검색의 새로운 프레임워크 모델(VvsS-Net)을 제안한다.

(b) 각 장면 그래프를 통해 시각적·의미적 유사도를 비교할 수 있는 새로운 접근 방식을 제안한다.

(c) MS-COCO [14]와 Visual-Genome [15] 데이터셋에 대해 다른 모델들에 비해 인간 동의 점수 측면에서 훨씬 좋은 성능을 보였다.

1.4 논문구성

본 논문은 다음과 같이 총 5장으로 구성된다. 제 2장에서는 본 연구에서 적용하고자 하는 장면 그래프와 그래프 합성곱 신경망(Graph Convolutional Network)를 소개하고, 이미지 검색 관련 선행 연구를 살펴본다. 제 3장에서는 본 연구에서 사용하는 라벨과 제안하는 기법들에 대해 설명한다. 제 4장에서는 실험에 사용한 데이터셋과 비교 데이터를 설명하고 실험 세팅 및 정량적·정성적 실험 결과를 통해 제안 기법의 효과를 증명한다. 마지막으로 제 5장에서는 결론과 한계점에 대해 논의하고 향후 연구 방향을 제시한다.

제 2 장 배경 이론 및 관련 연구

2.1 배경 이론

2.1.1 장면 그래프

장면 그래프(Scene graph)는 이미지의 복잡하고 자세한 의미들을 표현하기 위해 [11]에서 처음 제안된 개념이다. 이미지 상의 단순한 개체들(objects)에서 나아가, 개체들간의 관계(relationships), 개체들의 특성(attributes)와 같은 자세한 의미들(detailed semantics)을 표현한 그래프로 이미지의 장면을 표현하는 데이터 구조이다. 장면 그래프 그 자체로는 해당 이미지의 장면을 표현한 것이고 이미지에 대한 것은 아니지만, 이미지의 개체에 해당되는 부분인 경계 상자(bounding box)에 대응시킴(grounding)으로써 이미지와 장면 그래프 간에 연결 관계를 생성할 수 있다. 결과적으로 한 이미지에 대응되는 장면을 표현하는 장면 그래프와 장면 그래프의 개체들에 대응되는 이미지의 경계 상자 영역 라벨이 존재하게 된다.

장면 그래프의 특징은 개체, 관계, 특성을 표현하는 부분에 있어서 제약사항이 없다는 점이다. 개체는 사람, 장소, 물건, 다른 개체의 일부분이 될 수 있다. 관계는 개체들간의 관계가 될 수 있다면 관계 없으며, 어떠한 취하는 행동일수도 있고 위치 등이 될 수도 있다. 관계를 나타낼 때는 주로 <주어(subject), 술어(predicate), 객체(object)>의 튜플릿(tuplet)으로 방향이 있는 엣지(directed edge)로 표현된다. 개체의 특성은 모양, 색깔, 질감, 자세 등이 될 수 있다.

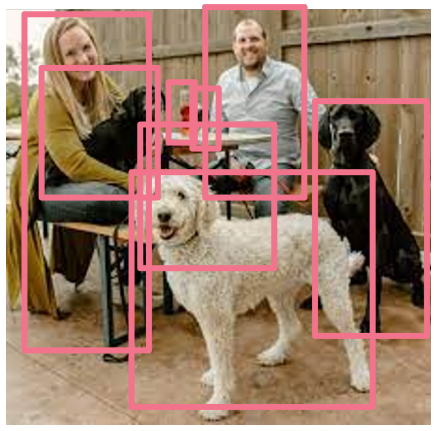
또한 장면 그래프는 구조화가 되어있기에 캡션과 같은 비구조화된 텍스트 형태의 자연어 라벨에 비해 재구조화할 필요가 없다는 장점이 있다. 또한 자연어 라벨은 구조화

후에 해당되는 개체들을 판별하고 이미지에 대응시키는 작업이 별도로 필요하다. 반면, 장면 그래프는 이미지에서 개체를 먼저 식별하고 해당 개체들로부터 관계와 특성들을 기술하면 된다는 장점이 있다.

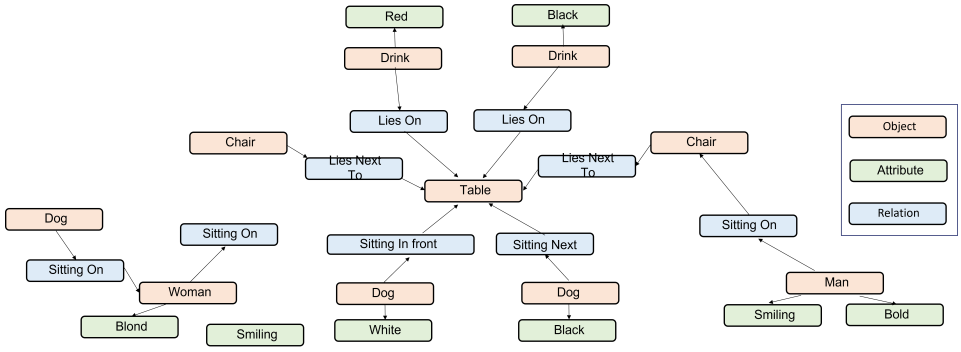
그림 2.1는 이미지와 장면 그래프의 예시로, 책상을 기준으로 책상 위에 음료들이 놓여 있고 남성과 개를 안고 있는 여성이 의자에 앉아있으며, 개 두 마리가 책상 앞에 앉아있는 이미지다. 개체들로는 '책상', '음료', '개', '여성', '남성', '의자'가 있다. 이 개체들간의 관계로 책상 위에 음료들이 '놓여 있다', 개를 '안고 있는' 여성, 의자에 '앉아 있는' 남성과 여성, 책상 '앞에 앉아 있는' 개들이 장면 그래프에서 관계를 표현하게 된다. 또한 각 개체들은 색깔과 같은 특징을 갖게 된다. 이렇게 장면 그래프가 구성되게 되고, 장면 그래프의 각 개체들이 이미지에 해당되는 영역에 경계 상자로 대응되게 된다.

본 연구에서는 이미지의 장면 그래프에서 이미지에서 각 개체를 나타내는 영역이 있다는 점과 단어 단위의 장면의 의미를 나타내는 그래프가 있다는 점을 이용해 장면 그래프로부터 시각적·의미적 특징을 추출하고자 한다. 이렇게 추출한 특징을 이용해 이미지 간 코사인 유사도(cosine similarity)를 통해 유사도를 계산하고자 한다. 그래프에서 특징 벡터(feature vector)를 학습시키기 위해 그래프 신경망(Graph Neural Network)가 활용이 되는데, 그 중 그래프 합성곱 신경망 구조를 채택하여 학습을 진행했다. 이에 대해서는 2.2에서 상세히 다루도록 한다.

(a) 원 이미지



(b) 이미지의 개체를 나타내는 경계 상자



(c) 이미지의 장면을 표현하는 장면 그래프

Figure 2.1: 이미지와 대응되는 경계 상자 및 장면 그래프 예시

2.1.2 그래프 합성곱 신경망

1) 그래프 모델

Convolutional Neural Network(CNN), Recurrent Neural Network(RNN) 등 다양한 종단간 학습(end-to-end learning)의 딥러닝 패러다임들이 등장하면서 유클리디언 공간의 데이터에 대해서는 성공적으로 특징을 추출할 수 있었다. 그러나, 데이터의 요소 간 상호의존성이 생기며 복잡한 관계를 데이터로 표현하면서 비-유클리디언 공간의 그래프 구조를 이용한 데이터들이 등장하기 시작했다 [16]. 대부분의 기존 딥러닝 모델들이 독립성을 가정한 모델들이 많았다면, 그래프에서는 이러한 독립성을 가정할 수 없다. 또한 합성곱연산(Convolution) 역시 일반 이미지에서는 합성곱의 대상이 되는 이웃 역시 정해져있으나, 그래프에서는 한 노드(node)가 몇 개의 엣지에 연결되어있는지는 정해지지 않은 불규칙성도 존재하기에 단순히 기존 딥러닝 모델을 적용하기에는 어려움이 있다.

이러한 그래프의 복잡한 특징을 다루기 위해, 기존 딥러닝 모델의 프레임워크에서 CNN, RNN 등을 그래프의 데이터에 접목하기 위한 연구가 활발히 진행되어왔다. 그래프 신경망 외에도 그래프 커널(Graph kernel) 기법들이 있었고, 이 역시 그래프나 노드를 매핑 함수를 통해 임베딩으로 만들 수 있었다. 그러나 그래프 신경망과 달리 이 매핑 함수가 사전 정의되어야 하며, GNN은 임베딩부터 분류 또는 유사도 추출까지 종단간 학습으로 학습할 수 있다는 장점이 있다. 또한 그래프의 유사성 문제에 한해 최적 수송 문제로 모델링하여 한 그래프를 다른 그래프로 표현하기 위해 드는 비용을 이용해 둘 간의 유사성을 계산하고자 하는 시도들도 있었다 [17, 18, 19, 20, 21]. 특히 이들 중 Gromov-Wasserstein을 이용한 방법들의 경우 [20, 21] 노드의 임베딩을 반영할 수 있기에 그래프의 연결 특징 뿐만 아니라 노드의 특성도 반영할 수 있다. 그러나 Gromov-Wasserstein의 경우 종단간 학습으로 학습 가능한 모델이 아니며, 2차

프로그래밍(quadratic programming)으로 이를 풀기위해 기존에 제안된 알고리즘들이 계산 비용이 높다는 단점이 있다 [22]. 그렇기에 본 연구에서는 그래프 신경망 모델으로 제한하여 그래프 간 유사도 계산을 위한 임베딩을 추출하고자 했다.

그래프 신경망 모델에는 CNN을 적용한 모델, RNN을 적용한 모델, 오토인코더를 적용한 모델 등 다양한 모델들이 있으나, 본 연구에서는 그래프 단위의 특징을 추출하는 것을 목표로 하기에 그래프 생성 분포를 학습하기 위한 오토인코더 모델이나 그래프의 숨겨진 패턴을 학습하기 위한 모델이 아닌 CNN을 적용한 그래프 합성곱 신경망 모델을 채택했다.

2) 그래프 합성곱 신경망

그래프 합성곱 신경망 모델은 [23]에서 처음 제안된 모델이다. 그래프 합성곱 신경망 모델은 각 노드의 특징을 해당 노드 뿐만이 아니라 주위의 이웃한 노드들의 특징을 통해 표현하는 것이 핵심으로, 여러 층의 그래프 합성곱 계층을 쌓음으로써 노드의 고차원 특징을 표현할 수 있다는 특징이 있다. 그래프 합성곱 신경망 모델은 그래프 전체의 특성을 추출할 때에는 그림 2.2와 같이 일반적으로 그래프 합성곱 계층(Grah Convolution Layers), 판독 계층(Readout Layers), 완전연결계층(Fully Connected Layers)의 총 3가지 요소로 구성되어 있다. 먼저 그래프를 그래프의 노드 특징 행렬과 인접 행렬로 표현한다. 이를 그래프 합성곱 계층에서 합성곱 연산을 통해 모든 노드에 대한 m차원 잠재 특성 행렬을 생성한다. 그래프 전체의 특성이 아닌 노드 특성 행렬만을 이용할 경우 이 계층의 결과물을 이용하게 된다. 그러나 이를 그래프 전체로 표현하는 하나의 벡터로 변환하는 과정을 판독 계층에서 거치고, 이 변환된 벡터에 대해 추가적인 연산을 하는 완전연결계층을 통과시켜 최종 그래프 특징을 도출하게 된다. 그래프에서 노드 특징을 추출하는 단계부터 최종 완전연결계층까지의 단계를 종단간 학습으로 학습시킬

수 있다는 장점이 있다.

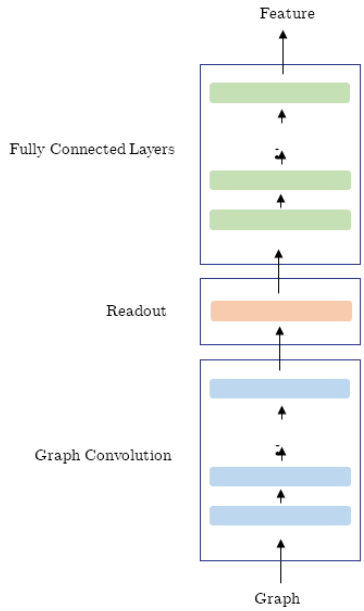


Figure 2.2: 그래프 합성곱 신경망 구조

그래프 합성곱 신경망에서는 엣지의 특성을 반영하지 않기에, 엣지의 특성이 필요한 경우 엣지를 노드로 만들고 연결된 노드들 사이에 엣지를 만든다. 총 노드의 개수가 N 개라고 할 때, 인접 행렬 A는 $N \times N$ 행렬로 표현이 가능하고, 각 노드의 특징을 d-차원 벡터에 대응시켜 노드 특성 행렬 X는 $N \times d$ 행렬로 표현이 가능하며, 이 때 그래프를 $G = (A, X)$로 정의할 수 있다.

그래프에 대한 합성곱 연산은 이 A와 X를 받아 새로운 잠재 노드 특성 행렬을 생성한다. 이 때 새로운 노드 특성은 이웃하는 노드의 특성을 합성곱 연산을 통해 반영하게 된다. 이미지에서 합성곱 연산이 인접해있는 지역의 정보를 모으듯이, 그래프에서는

연결되어있는 노드의 특징을 모음으로써 지역적인 정보를 취합해 잠재 특성을 고도화해나가는 과정이다. 가장 기본적인 합성곱 연산은 다음과 같이 정의할 수 있다.

$$H = \psi(A, X) = \sigma(AXW) \tag{2.1}$$

W는 d 차원 특성을 m 차원으로 매핑시켜주는 $d \times m$ 의 학습 가능한 행렬로 모든 노드들에 대해 해당 가중치를 공유한다는 점에서 기존 이미지 합성곱 연산과 유사하다. 인접행렬, 노드 특성 행렬, 가중치 곱한 후 비선형 함수 $\sigma(\cdot)$ 를 통과시켜 그 다음 층위의 잠재 노드 특성 $N \times m$ 차원의 행렬 H를 계산한다.

이렇게 기존 합성곱 연산과 유사한 단순한 방식의 경우 다음과 같은 한계점들을 갖고 있다.

- 인접 행렬에는 이웃 노드와의 연결 정보만을 담고 있기에 연산 과정에서 노드 자신의 특성은 합성곱에서 고려되지 않는다.
- 인접 행렬이 정규화되어있지 않기에, 특성의 크기가 불안정할 수 있다. 이웃 노드가 많은 노드의 경우 그라디언트 폭팔(exploding gradient), 적은 노드의 경우 그라디언트 소실(vanishing gradient)의 문제 등이 발생할 수 있다.

두 가지 한계점을 극복하기 위해 다음과 같이 self-loop을 추가하고 차수를 이용해 인접행렬을 정규화한 행렬을 사용하여 그래프 합성곱 연산을 하게 된다.

$$\hat{A} = A + I \tag{2.2}$$

$$\psi(\hat{A}, X) = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{1/2}XW) \tag{2.3}$$

그래프 합성곱 연산 계층의 각 층에서는 이전 잠재 노드 특성을 받아 2.3와 거의 동일한

연산을 각 층마다 2.4로 수행하여 노드의 최종 잠재 특성 행렬을 추출하게 된다. $\hat{A}$는 인접 행렬에 Identity matrix를 더함으로써 self-loop를 추가한 그래프의 인접 행렬이고, $\hat{D}$는 $\hat{A}$의 차수 행렬로 $\hat{A}$는 $\hat{D}^{-1/2}\hat{A}\hat{D}^{1/2}$와 같이 차수 행렬 $\hat{D}$를 이용해 정규화시킨다. 정규화한 행렬을 2.4의 A 대신 사용한 수식과 동일하다.

$$H^{(k)} = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{1/2}H^{(k-1)}W^{(k)}) \tag{2.4}$$

이후 판독 계층을 통해 그래프 합성곱 계층을 통해 추출한 노드의 최종 잠재 특성 행렬을 그래프를 표현하는 하나의 벡터 특성으로 변환을 진행한다. 풀링과 유사한 개념으로, 이미지 연산에서는 주로 가장 강한 특성을 담아내기 위해 최대 풀링(max pooling)을 사용하지만 그래프 연산에서는 주로 평균 풀링(average pooling)을 사용한다. 평균 풀링을 이용한 판독 계층 이후 그래프의 특성은 다음과 같이 정의되며, 그래프 합성곱 계층의 마지막 결과가 $N \times d^{(l)}$ 차원의 행렬이라 했을 때 노드들에 대해 평균낸 벡터로 $d^{(l)}$ 차원의 벡터가 된다.

$$\mu(G) = \frac{1}{N}\sum_{i=1}^{N} H_i^{(l)} \tag{2.5}$$

완전 연결 계층은 판독 계층을 통해 추출한 그래프의 특성에 추가적인 연산을 위한 계층으로 주로 그래프 단위의 분류 등의 모델 학습에 추가로 사용된다. 그러나 본 연구에서는 그래프 간에 유사도를 비교하기 위해 그래프 합성곱 신경망 구조를 채택한 것으로, 그래프 전체의 특성을 담아낸 2.5을 이용해 코사인 유사도를 사용하여 그래프 간 유사도를 비교한다. 그렇기에 별도의 연산 작업인 완전 연결 계층의 경우 생략하여 모델 학습을 진행했다.

2.2 관련 연구

본 연구에서는 이미지의 장면 그래프로부터 그래프 합성곱 신경망을 이용해 시각적·의미적 특징을 각각 도출하는 모델을 학습하고, 각 특징들을 사용자의 의도에 맞게 비율을 조정하여 이미지 검색 기법을 제안하고자 한다. 이를 위해 이미지에서 특징을 추출하는 연구들과 시각적·의미적 유사도에 집중한 이미지 검색에 대한 관련 연구들을 살펴보도록 한다.

2.2.1 이미지 특징 추출 연구

CNN으로부터 추출된 이미지 특징은 다양한 컴퓨터 비전 분야에서 좋은 성능을 보였다. 초기 연구는 이미지 분류 작업을 위한 특징을 이미지 단위의 라벨을 이용해 학습했다 [24, 25]. 이후 앵커(anchor) 이미지에 대해 두 이미지 중 어떤 것이 더 유사한지에 대한 정보를 가진 트리플렛(triplet)을 이용한 metric learning으로 적은 데이터 요구 사항으로도 이미지의 특징을 학습 가능한 모델들이 제시되었다 [26].

이미지에 대한 시각적·의미적 특징을 모두 함께 학습하기 위한 클래스 라벨의 단어 임베딩을 활용한 연구 [27, 28], 분류를 위한 클래스 구조를 이용하는 연구 [29], 클래스 계층을 WordNet 온톨로지로 활용한 연구 [30, 31]들이 있었다. 이러한 방법론들은 이미지의 의미를 이미지 전체를 표현하는 클래스 단어 하나에 의존적이다. 그렇기에 이미지의 단순한 의미는 담을 수 있지만, 여러 개체와 관계가 있는 복잡한 장면의 이미지에 확장하기엔 한계가 있었다. 보다 최근 연구들에서는 캡션과 같은 텍스트와 이미지를 동시에 이용하는 멀티모달 학습으로 랭킹 수식 [32, 33] 혹은 similarity 네트워크 [34]로 시각적·의미적 특징을 학습하고자 했다.

2.2.2 이미지 검색

이미지 검색은 표준 벤치마크 데이터셋에서 정확히 같은 객체를 찾는 객체 위주 검색의 연구가 주로 진행되어 왔다 [2, 3, 4, 5, 35]. 객체 위주의 검색 외에는 같은 카테고리 라벨 [36, 37]혹은 태그 [38]를 갖는 이미지 검색 연구가 있었다. 그러나 이러한 연구들은 장면의 의미를 조악하게 이해한다는 한계가 있었다. [39]에서는 추상적인 장면을 담은 합성 데이터셋을 이용해서 이미지에 대한 자세한 의미들을 반영 가능할 때 이미지 검색 성능이 크게 향상됨을 보였다.

이미지의 장면을 명시적으로 모델링하기 위한 연구들로 개체의 특성들을 이용하는 연구 [40], 겹치는 개체 수(object co-occurences)를 이용한 연구 [41], 개체 간의 관계를 이용한 연구들 [42, 43, 44]이 있었다. 장면 안의 개체 간에 상호 작용이 복잡해질 수 있기에 단순한 동시 발생, 쌍으로 비교한 관계들에서 나아가 시각적 특징 대신 명시적인 장면 그래프를 비교하는 방식이 제안되었다 [3]. 이 연구의 경우 사용자가 쿼리를 장면 그래프의 형식으로 주어야 한다는 단점이 있기에 실제로 적용하기 힘들다는 한계점이 있다. 또한, 해당 연구는 장면의 복잡한 관계를 단어 단위의 장면 그래프만으로 모델링하기에 시각적 특징을 활용하지 않는다는 단점이 있다. 그러나 그래프 단위의 비교가 유의미한 이미지를 검색 결과로 반환할 수 있다는 점과 시각적 특징을 명시적으로 이용하지 않고 단어로만 구성된 장면 그래프 그 자체로도 유의미한 이미지 검색을 할 수 있음을 보였다. 그 외에 자세한 의미를 명시적으로 모델링하기 위한 연구들로는 이미지 캡션을 이용한 연구들이 있었다 [45, 46, 47]. 캡션을 이용해 텍스트로 이미지를 검색하기 위해 단순한 전역 특징을 사용하거나, ImageNet을 이용해 사전학습된 특징을 사용하거나 개체 탐지기 혹은 장면 분류기로부터 얻은 복잡한 특징을 사용했으나 이 특징들은 모두 검색을 위해 학습된 특징이 아니라는 한계가 있다.

기존 연구들을 통해 이미지 검색에서 객체 동일성을 중시하는 시각적 유사도 기반

의 검색도 CNN에 기반해 활발히 연구가 되고 있고, 복잡한 장면의 의미를 담기 위한 자세한 의미들을 중시하는 의미적 유사도 기반의 검색 역시 개체, 캡션, 장면 그래프 등에 기반해 활발히 연구가 진행이 되고 있음을 확인할 수 있다. 그러나 이들을 상황에 맞게 동시에 이용하거나 제어하려는 연구는 부족한 상황이다.

제 3 장 제안 기법

본 연구에서 제안하는 기법을 이하 VvsS-Net(Visual versus Semantic-Net)으로 칭한다. 시각적 유사도와 의미적 유사도 간의 비율을 조정하여 이미지 간 유사도를 계산하여 검색하는 프레임워크이다. 먼저 단일 쿼리 이미지를 받아서 이미지로부터 장면 그래프를 생성한다. 그 후 후보 데이터셋의 장면 그래프들과 비교 작업을 수행하게 된다. 이 때 장면 그래프 간의 시각적 유사도와 의미적 유사도는 각각 그래프의 시각적 임베딩과 그래프의 의미적 임베딩으로부터 계산하게 된다. 각 임베딩은 대응되는 대리 관련도를 이용해 그래프 합성곱 신경망을 학습시킨 결과물이다. 마지막으로 VvsS-Net은 시각적 및 의미적 유사도를 반영한 전체 유사도를 기준으로 이미지를 검색한다. 이 때 시각적·의미적 유사도의 비율을 조정할 수 있다.

3.1에서는 각 임베딩을 학습시키기 위해 이미지 간 대리 관련도를 추출한 방법에 대해 제시하고, 3.2에서는 장면 그래프에서 각각 어떤 방법으로 시각적·의미적 그래프를 도출하였고 그로부터 그래프 임베딩을 학습시킨 방법과 이를 바탕으로 검색 방법에 대해 제시한다.

3.1 이미지 대리 관련도 추출

그래프 임베딩을 학습시키기 위해서 라벨로서 두 그래프가 얼마나 유사한지에 대한 점수 혹은 쿼리 그래프에 대해 positive한 그래프와 negative한 그래프 라벨이 있어야 한다. 그러나 이러한 데이터는 없기에 두 그래프에 대응되는 이미지들 간에 다른 방식으로 구한 유사도를 약한 감독 학습(weak supervised trainin) 라벨로 사용하기로 한다. 이미지의 시각적 특징에 집중한 대리 관련도와 의미적 특징에 집중한 대리 관련도 각각 필요하다.

3.1.1 시각적 대리 관련도

시각적 대리 관련도 $S_v(I_i, I_j)$를 도출하기 위해 이미지 I_i와 I_j의 시각적 특징 $\phi(I_i)$와 $\phi(I_j)$ 벡터를 사용한다. 각 이미지의 시각적 특징 벡터는 이미지의 사전 학습된 ResNet-152[24]로 얻은 이미지 특징을 이용한다. ResNet은 vanilla CNN에 잔차(residual)을 추가한 모델로 사용한 ResNet-152는 총 152층의 깊은 구조로 되어있다. [3]에 따르면 ResNet이 기존의 얕은 VGG[48]에 비해 이미지 검색 성능에서 더 좋은 성능을 보임을 확인했기에 사전 학습된 ResNet-152 모델을 채택했다. 이를 통해 $\phi(I_i)$를 I_i의 ResNet feature를 단위구에 사영시킨 단위 벡터로 정의한다. 이를 이용해 시각적 대리 관련도를 3.1와 같이 $\phi(I_i)$ 와 $\phi(I_j)$의 내적으로 정의할 수 있다.

$$S_v(I_i, I_j) = \phi(I_i) \cdot \phi(I_j) \tag{3.1}$$

3.1.2 의미적 대리 관련도

의미적 대리 관련도 $S_s(I_i, I_j)$를 도출하기 위해 각 이미지 I_i, I_j에 대응되는 사람이 라벨링한 캡션 c_i와 c_j을 사용한다. 캡션은 이미지의 의미 정보를 표현하고 있는 텍스트로 주로 문장으로 주어진다. [9]에서 캡션을 이미지 검색에 사용했듯이, 캡션이 이미지 상의 세세한 정보는 담아내지 못한다고 하더라도 주요한 개체, 속성, 관계 정보를 가지고 있을 가능성이 높기에 이 캡션 사이의 유사도는 두 이미지 간의 맥락적 유사도를 담아낼 수 있다고 할 수 있다. 각 이미지의 의미적 특징 벡터는 사전 학습된 Sentence-BERT(SBERT) [49] 을 캡션에 적용한 결과를 이용한다. BERT [50]는 자연어처리에 특화된 트랜스포머 모델 중 하나로 SBERT는 이를 문장 단위에 적용했을 때의 성능을 높인 모델이다. 문장 간의 쌍 분류 작업과 문장 간의 쌍 회귀 작업으로 학습시킨 모델로, 문장 간의 관계 분류(Natural Language Inferencing), 문장 간의 유사도 계산(Semantic Textual Similarity) 등으로 학습시킨 결과이므로 문장 간의 유사도를 계산하기에 적합한 모델이라고 판단하여 채택했다. 또한 SBERT는 Siamese Network를 이용해 학습시킨 결과기에 코사인 유사도를 사용하기에 적합하다. 이를 통해 $\psi(I_i)$를 I_i의 대응되는 캡션 c_i의 SBERT feature을 단위구에 사영시킨 단위 벡터로 정의한다. 이를 이용해 의미적 대리 관련도를 3.2와 같이 $\psi(I_i)$와 $\psi(I_j)$의 내적으로 정의할 수 있다.

$$S_s(I_i, I_j) = \psi(c_i) \cdot \psi(c_j) \tag{3.2}$$

한 이미지에 여러 개의 캡션 라벨이 존재할 경우, 모든 캡션 쌍에 대한 유사도 평균을 사용했다.

3.2 장면 그래프를 통한 이미지 간 유사도 계산

그림 3.1은 본 연구에서 제안하는 VvsS-Net 학습 방식을 도식화한 그림이다. 이미지로부터 장면 그래프를 생성하고, 장면 그래프의 각 노드 당 경계 상자로부터 얻은 시각적 특징과 노드를 표현하는 단어로부터 얻은 의미적 특징으로 노드 특성 행렬을 구성한다. 이를 그래프 합성곱 신경망을 이용해 그래프 임베딩을 도출하고 이들의 코사인 유사도를 대리 관련도를 이용해서 그래프 합성곱 신경망의 파라미터 값들의 학습을 진행한다.

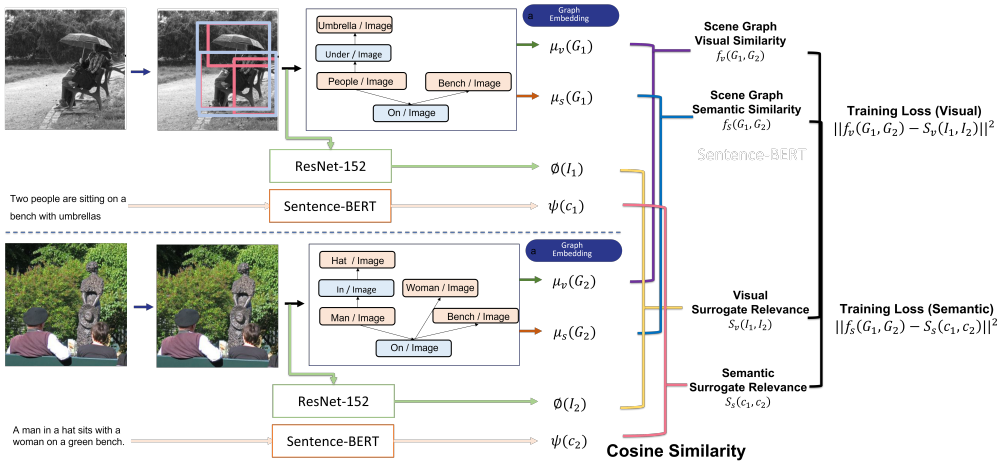


Figure 3.1: 본 연구에서 제안하는 VvsS-Net 학습 방식

본 연구에서는 개체의 특성에 대한 노드들은 제외를 하고 노드를 구성했다. 개체의 특성은 주로 색깔, 질감, 모양 등에 관한 것으로 의미 정보가 아닌 시각 정보에 해당되는 경우가 많기에, 노드 구성 시에 특성 정보를 넣을 경우 시각 정보에 치우쳐질 가능성이 있다. 그렇기에 이러한 특성 노드는 제외함으로써 시각적 유사도와 의미적 유사도 간의 제어에 중립성을 더해줄 수 있다. VvsS-Net에서 사용하는 그래프의 노드는 최종적으로

개체들과 이들 간의 관계를 나타내는 노드들이다. 시각적 특징 추출 그래프와 의미적 특징 추출 그래프 모두 대응되는 노드와 인접 행렬은 같으나 각 노드의 초기 특성 벡터가 다른 그래프 구조이다.

3.2.1 시각적 특징 추출 그래프

시각적 특징을 추출하기 위한 그래프의 노드 특성으로는 각 노드의 시각적 특징을 담아내야 한다. 노드 특성 역시 종단간 학습으로 임의로 초기화된 벡터에서 학습시켜 나갈 수 있으나, 학습 속도 개선 및 초기 방향성 설정을 위해 사전 학습된 ResNet-152 모델을 사용하여 초기화한 후 파라미터 값을 갱신해나가는 방식을 채택했다. 개체 노드의 경우 이미지에 해당되는 경계 상자를 이용해 해당되는 영역을 잘라서 이용하는 방식을 이용했고, 관계 노드의 경우 별도의 경계 상자가 없기에 관계에 연결된 주어와 객체의 경계 상자를 합한 영역을 이용했다. 이 때, 직사각형의 이미지 패치를 유지하기 위해 경계 상자를 합할 시에는 왼쪽 위(top-left)와 오른쪽 아래(bottom-right)를 이용해 관계 노드의 경계 상자를 설정했다. 그 후 이 영역의 ResNet feature을 그래프의 초기 노드 시각적 특징으로 사용했다.

3.2.2 의미적 특징 추출 그래프

의미적 특징을 위한 노드 특성 역시 학습시킬 수 있는 파라미터지만, 사전 학습된 Glove [51] 단어 임베딩을 사용하여 초기화한 후 파라미터 값을 갱신해나가는 방식을 채택했다. 장면 그래프의 모든 노드들은 해당되는 단어들을 사용했고, 해당 단어들의 Glove 임베딩을 그래프의 초기 노드 의미적 특징으로 사용했다.

3.2.3 모델 학습

그래프 간 유사도는 대리 관련도 계산과 비슷한 방식으로 그래프 임베딩을 단위 벡터화 후 내적한 값을 사용했다. 이미지 I_i에 대응되는 장면 그래프 G_i의 시각적 임베딩을 $\mu_v(G_i)$, 의미적 임베딩을 $\mu_s(G_i)$라고 할 때, 시각적 유사도 $f_v(G_i, G_j)$와 의미적 유사도 $f_s(G_i, G_j)$는 각각 다음과 같이 내적으로 계산된다.

$$f_v(G_i, G_j) = \mu_v(G_i) \cdot \mu_v(G_j) \tag{3.3}$$

$$f_s(G_i, G_j) = \mu_s(G_i) \cdot \mu_s(G_j) \tag{3.4}$$

이들 각각을 대응되는 대리 관련도 $S_v(I_i, I_j)$, $S_s(I_i, I_j)$와의 평균 제곱 오차(mean squared error)를 최소화하도록 L2 손실함수를 사용하여 그래프의 노드 특성과 그래프 합성곱 신경망 모델의 파라미터를 학습한다.

$$L_v(I_i, I_j) = ||f_v(G_i, G_j) - S_v(I_i, I_j)||^2 \tag{3.5}$$

$$L_s(I_i, I_j) = ||f_s(G_i, G_j) - S_s(I_i, I_j)||^2 \tag{3.6}$$

3.2.4 시각적·의미적 유사도의 비율을 반영한 이미지 간 유사도 추론

추론을 위한 최종 이미지 간 유사도는 다음 3.7과 같이 정의할 수 있다.

$$f_{vs}(G_i, G_j) = \lambda f_v(G_i, G_j) + (1 - \lambda) f_s(G_i, G_j) \quad , \quad 0 \leq \lambda \leq 1 \tag{3.7}$$

G는 이미지로부터 생성된 장면 그래프 혹은 실제 장면 그래프이고, f_v와 f_s는 각각 장면 그래프로부터 도출한 시각적·의미적 유사도이다. λ가 이들 간의 비율을 제어하는 파라미터(VS-gain)이다. $\lambda = 1$일 때는 이미지의 시각적 특징에 중점을 두고 유사도를 계산할 수 있고, $\lambda = 0$일 때는 이미지의 의미적 특징을 강조해 유사도를 계산하게 된다.

제 4 장 실험 결과

4.1 데이터셋

4.1.1 VG-COCO

본 연구에서는 이미지에 대해 캡션과 장면 그래프가 모두 있는 데이터셋을 사용하기 위해 공개 데이터셋인 Visual-Genome [15] 데이터셋과 MS-COCO [14] 데이터셋을 함께 사용해 데이터셋을 구축했다.

Visual-Genome [15] 데이터셋은 이미지의 구조화된 컨셉을 언어에 연결하려는 지식 베이스 데이터셋이다. 총 108,077개의 이미지로 되어있으며, 각 이미지에 대해 장면 그래프와 Visual Queteion Answers(VQA)에 대한 데이터를 갖고 있다. 이 중 장면 그래프 정보만을 이용하고, 각 장면 그래프는 개체, 관계, 특성으로 구성되어있다. 각 개체는 경계 상자 정보가 주어지고, 모든 노드에 대한 단어 정보는 Wordnet synset [52]으로 매핑되어있다. 현재 컴퓨터 비전 분야에서 벤치마크 데이터셋 중 하나로 [53]에서 Visual-Geonme 데이터셋을 일부 수정한 버전을 본 연구의 데이터셋으로 사용했고, 학습-검증 데이터셋 분할 역시 이 데이터셋의 방식을 이용했다.

MS-COCO [14] 데이터셋은 이미지에 대한 개체 탐지(Object detection) 및 segmentation 데이터, 캡션으로 이루어진 데이터셋이다. 총 330,000개의 이미지로 되어있으며, 각 이미지에 대해 사람이 작성한 5개의 캡션이 있다.

Visual-Genome 데이터셋과 MS-COCO 데이터셋의 교집합(VG-COCO)을 본 연구의 데이터셋으로 사용함으로써 각 이미지에 대한 장면 그래프 및 개체의 경계 상자와 5개의 캡션을 데이터로 구축할 수 있었다. 빈 장면 그래프를 가진 이미지들을 제외하

Figure 4.1: 학습 데이터 이미지 예시

고 총 48,220개의 데이터로 각각 학습 데이터로 35,017개, 검증 데이터로 13,202개의 데이터를 이용했다.

그림 4.1은 학습 데이터 중 하나로 강가 근처의 벤치에 여성과 남성이 기대있는 이미지이며 Visual Genome id 61534, MS-COCO id 460059에 해당되는 이미지다. 그림 4.2이 이 데이터의 Visual Genome 라벨로 학습에 사용할 경계 상자 라벨과 장면 그래프 라벨만을 이용했다. 해당 장면 그래프에서 특성 노드 (보라색 라벨)을 제외하고 학습에 활용했다. 그림 4.3이 이 데이터의 MS-COCO 라벨로 개체에 대한 segmentation 라벨도 존재하지만 캡션 데이터만을 학습에 활용했다. 캡션 라벨은 다음과 같다.

- man and woman sitting on a bench near a body of water.
- two people sitting on a wooden bench looking at the camera.
- a man and a woman are sitting on a park bench.
- two adults seated on a wooden bench near a river.
- a woman is wearing a yellow jacket and a man in a white shirt

(a) 경계 상자 라벨

(b) 장면 그래프 라벨 [15]

Figure 4.2: 4.1의 Visual Genome 라벨

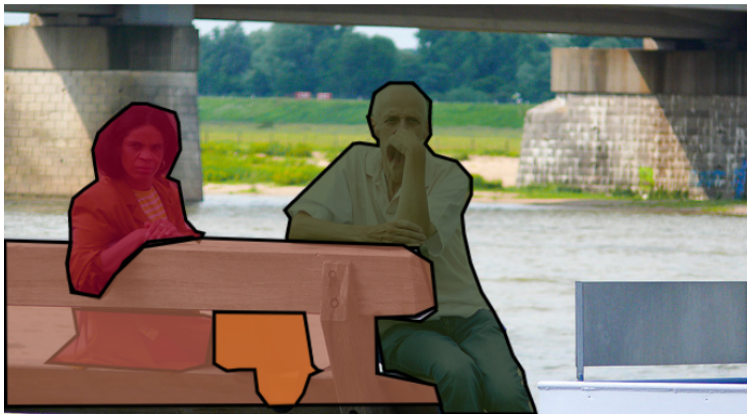


Figure 4.3: 4.1의 MS-COCO segmentation 라벨

VG-COCO 테스트 데이터셋에서 이미지 검색을 위해 테스트셋에서 1000개의 쿼리 이미지 셋을 임의로 선정했고, 각 쿼리 이미지에 대해 나머지 13,202개의 테스트 이미지들 중 이미지 검색을 진행했다.

4.1.2 캡션 및 장면 그래프 생성

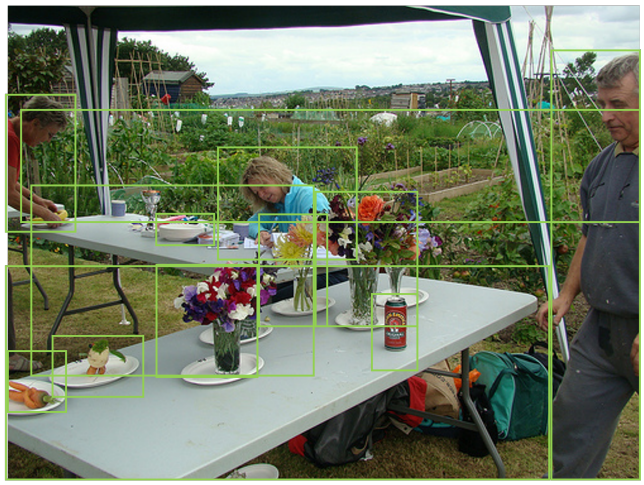
VG-COCO의 캡션과 장면 그래프 데이터를 모두 이용하기로 했으나, 현실 또는 평가 단계에서 모든 이미지들에 대해 사람이 작성한 캡션과 장면 그래프 데이터를 얻는 것은 실용성이 부족하다. 그렇기에 각 이미지에 대해서 모델을 이용해 생성한 캡션과 장면 그래프 역시 사용하도록 했다.

사람이 라벨링한 캡션이 아닌 기계에 의한 이미지에 대한 캡션 생성으로는 Flickr30k [54] 데이터셋으로 사전 학습된 soft-attention 모델 [55]을 사용하여 이미지들에 대한 캡션을 생성했다.

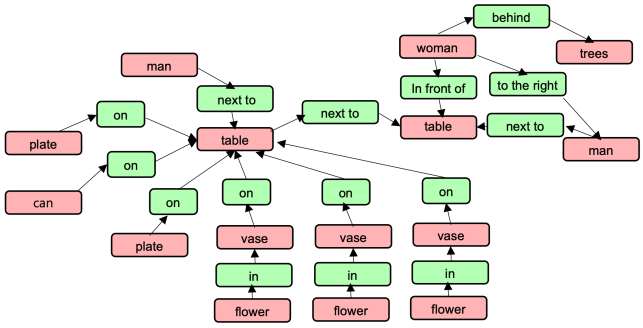
Figure 4.4: 테스트 데이터 이미지 예시

사람이 라벨링한 장면 그래프가 아닌 기계에 의한 이미지에 대한 장면 그래프 생성으로는 전통적인 장면 그래프 생성 방식을 이용했다. [56]와 같이 이미지로부터 Faster R-CNN 모델을 이용해 개체들을 탐지하고 confidence 0.3을 기준으로 최대 100개의 개체만을 노드로 남긴 뒤, 해당 경계 상자로부터 얻은 이미지로부터 사전 학습된 ResNet-101 feature을 통해 개체의 이름과 특성들을 예측했다. 개체 간의 관계 라벨을 추출하기 위해 총 309 종류의 관계 라벨이 있는 GQA 데이터셋 [57]으로부터 구성된 빈도 기반의 사전 지식(frequency prior knowledge)를 이용했다. 탐지된 개체들의 모든 쌍에 대해 confidence 0.2를 기준으로 관계를 예측했다. 이러한 장면 그래프 생성 알고리즘이 [58, 53, 59]에 비해 단순한 알고리즘이지만 성능에 있어서 더 우수했기에 해당 방식을 채택해서 장면 그래프를 생성했다.

그림 4.4는 테스트 데이터 중 하나로 야외에 꽃이 올려진 테이블이 있고, 테이블 주위에 사람들이 서있는 이미지이며 Visual Genome id 2330191, MS-COCO id 356916에 해당되는 이미지이다. 생성된 캡션은 "a group of people sitting around a table"이고, 그림 4.5는 이 이미지로부터 특성 노드를 제외한 생성된 장면 그래프와 해당되는 경계 상자이다.



(a) 경계 상자



(b) 생성된 장면 그래프

Figure 4.5: 4.4의 생성된 경계 상자 및 장면 그래프

(a) 쿼리 이미지

(b) 후보 이미지 1

(c) 후보 이미지 2

Figure 4.6: 이미지 트리플렛 예시

4.1.3 인간 동의 점수

대리 관련도를 라벨로 이용한 결과는 이미지 검색 알고리즘 성능이 얼마나 인간의 평가와 유사한지를 검증하기 어렵다. 그렇기에 사람들로부터 이미지 간 유사도에 대한 라벨을 수집했다. 사람들 간에 생각하는 유사도 수치가 다를 것이기에 정확한 수치를 수집하기보다는 쿼리 이미지에 대해 후보 이미지 두 개 중 어떤 것이 더 가까운지를 평가받는 것이 더 유의미한 라벨이 될 것이다.

그림 4.6은 라벨러에게 제시한 이미지 트리플렛 예시로, 아래의 총 네 가지 선택지 중 하나를 답하게 된다.

(a) 후보 이미지 1이 쿼리와 더 유사하다

(b) 후보 이미지 2가 쿼리와 더 유사하다

(c) 두 이미지가 우열을 가릴 수 없게 쿼리와 유사하다

(d) 두 이미지 모두 쿼리와 유사하지 않다

29명의 라벨러로부터 1,752개의 이미지 트리플렛에 대해 10,712개의 데이터를 수집했다. 한 사람의 평가로 인해 치우쳐지지 않도록 한 트리플렛당 평균 6.1개의 응답을 수집했다.

쿼리 이미지는 4.1에서 정의한 쿼리 데이터셋에서 한 개를 임의로 선택했고, 아래 두 가지 조건에 해당되는 테스트 데이터셋 이미지들 중 임의로 두 개를 선택했다.

(a) ResNet-152 코사인 유사도를 이용해 랭킹한 결과, 100위 안의 이미지

(b) 대리 관련도 차이가 0.1 이상 나는 이미지

모델의 성능을 평가하기 위한 인간 동의 점수는 [9]의 방식으로 도출했다. 각 트리플렛마다 인간이 작성한 라벨과 같은 결정을 하는 평균 비율을 인간 동의 점수(human agreement score)로 사용했다. 수식으로 표현하자면, s_1, s_2가 각각 후보 1, 2를 선택한 사람 수(a, b), s_3이 두 개가 다 매우 비슷하다(c), s_4가 둘 중 어느것도 비슷하지 않다(d)고 대답한 사람 수일 때, 알고리즘이 후보 이미지 1이 더 가깝다고 판단한 경우 해당 트리플렛의 인간 동의 점수는 $w_i/(s_1 + s_2 + s_3 + s_4)$ where $w_i = (s_1 + 0.5s_3)\mathbb{1}_{i=1} + (s_2 + 0.5s_3)\mathbb{1}_{i=2} + (0.5s_1 + 0.5s_2 + s_3)\mathbb{1}_{i=3}$로 정의할 수 있다. 두 이미지 중 랜덤하게 고를 경우 평균 동의 점수는 0.472(표준편차 0.01), 사람들 간에 평균 동의 점수는 0.727(표준편차 0.05)이다.

이 때 이전 연구 [10]의 경우 알고리즘이 무조건 두 이미지 중 선택을 해야했다면, 본 연구에서는 사람의 판단과 유사하게 역치를 두고 둘 다 유사하다와 둘 다 유사하지 않다 중에 선택이 가능하도록 변경했다.

4.2 실험 세팅

4.2.1 2단계 이미지 검색

본 연구의 실험은 2단계로 이미지 검색을 진행했다. 비슷한 이미지를 대강 검색한 후, 그들 사이에 유사도를 사용하여 랭킹을 다시 진행해서 검색 결과를 반환한다. 이로 인해 계산 비용을 줄일 수 있고, 좋은 후보 이미지셋을 만들 수 있다.

쿼리 이미지에 대해 사전 학습된 ResNet-152 feature을 이용해서 가까운 총 100개의 이미지를 검색한 후 각 모델을 이용해 유사도를 재계산하여 결과를 반환했다.

4.2.2 학습 세팅

실험 시 학습 세팅에 대해 기술하면, 시각적 모델과 의미적 모델 모두에 대해 초기 학습률(learning rate)이 0.0001인 Adam optimizer를 사용하고, 매 에폭(epoch)마다 학습률에 0.9를 곱했다. 배치(batch) 크기는 16으로 설정하였고 총 20 에폭 학습을 진행하였다. 학습 시에 이미지 쌍을 오버샘플링 기법을 사용해서 샘플링했다. 한 앵커 이미지에 대해 총 세 개의 이미지를 샘플링했는데, 첫 번째는 시각적 대리 관련도를 기준으로 100개의 유사한 샘플들 중에 샘플링, 두 번째는 의미적 대리 관련도를 기준으로 100개의 유사한 샘플들 중에 샘플링, 세 번째는 그 외의 이미지들 중 하나를 샘플링했다. 이러한 오버샘플링 기법을 적용함으로써 유사한 이미지에 대해 더 학습하기 적합하도록 했다.

그래프 합성곱 신경망의 그래프 합성곱 계층은 각각 1024개의 히든 뉴런(hidden neuron)로 구성된 3층을 이용했다. 그 후, 최종 임베딩 도출을 위해 판독 계층으로는 최대 풀링에 비해 평균 풀링이 더 좋은 성능을 보임을 기존 연구 [10]에서 확인한 결과, 평균 풀링을 사용하고 코사인 유사도 도출을 위해 단위 벡터로 크기 변환을 진행했다.

4.3 실험 결과

4.3.1 정량적 실험 결과

본 연구의 결과를 시각적·의미적 대리 관련도에 사용된 모델 각각을 정답 라벨로 간주했을 때 계산한 ndCG(normalized discounted cumulative gain) 점수를 통해 각 그래프 모델이 잘 학습되어있는지를 검증했다. 또한, 시각적·의미적 유사도의 비율을 제어한 VvsS-Net 모델이 인간 동의 점수에 있어서 다른 베이스라인 모델들에 비해 더 좋은 성능을 보임을 통해 두 측면 모두 고려한 검색이 이미지 검색 성능에 유의미한 향상을 보임을 검증했다.

1) 베이스라인 모델

사전 학습된 ResNet-152 모델 사전 학습된 ResNet-152 모델을 통해 추출한 이미지 전체 임베딩의 코사인 유사도를 이용해 검색한 결과. 시각적 대리 관련도에 사용된 모델이다.

개체 수(Object Count) 장면 그래프의 개체 노드만을 이용해 개체 수 노드 벡터의 코사인 유사도를 이용해 검색한 결과

캡션 SBERT SBERT 모델을 통해 추출한 캡션 임베딩의 코사인 유사도를 이용해 검색한 결과. 정답 라벨을 이용한 결과와 생성된 라벨을 이용한 결과 모두를 비교한다.

ResNet feature와 SBERT feature의 내삽 결과 시각적·의미적 유사도를 동시에 사용하는 방법으로 시각적 ResNet feature와 의미적 SBERT feature의 내삽을 이용해 본 연구와 비슷한 방식으로 내삽한 모델이다. 본 연구와의 차이점은 이들은 각각 이미지 분류를 위해 학습된 모델, 문장 임베딩을 위해 학습된 모델로 다른 층위로 학습된 벡터이다. 그러나 본 연구에서 제안하는 임베딩은 같은 그래프 층위에서 학습된 결과라는 차이점이 있다. 이 베이스라인 모델과의 비교를 통해 나이브한 접근 방식보다

그래프 단위의 비교가 유의미했다는 것을 검증하고자 했다.

2) 각 대리관련도를 이용한 nDCG 결과

nDCG [60]는 추천 시스템에서 랭킹 추천 분야에 많이 쓰이는 평가 지표로 1에 가까울수록 정답에 가까운 지표이다. 그래프 임베딩이 대리 관련도를 이용해 잘 학습되었는지를 검증하기 위한 지표로, 대리 관련도를 정답 라벨로 간주하고 알고리즘의 랭킹 결과를 평가한 결과이다. 시각적 대리 관련도를 이용한 nDCG 결과는 표 4.1, 의미적 대리 관련도를 이용한 nDCG 결과는 표 4.2와 같다. 장면 그래프는 모두 사람이 라벨링한 VG-COCO의 라벨 장면 그래프를 이용했다. VvsS-Visual은 그래프 시각적 임베딩, VvsS-Semantic은 그래프 의미적 임베딩, VvsS-half는 그래프 시각적 임베딩과 의미적 임베딩을 0.5와 0.5로 내삽한 결과, VvsS-Best는 인간 동의 점수에서 가장 높은 점수를 얻은 비율로 내삽한 결과, Cap(Gen)은 생성된 캡션을 이용한 결과이다.

표 4.1로부터 자기 자신 모델을 제외하고 그래프의 시각적 임베딩 모델이 다른 모델들에 비해 높은 nDCG 점수를 보임을 확인할 수 있다. 실험이 ResNet feature을 이용해 후보를 만든 후 다시 랭킹하는 접근 방식을 취했기에, 다른 베이스라인 모델들의 nDCG 점수 역시 높은 편이지만, VvsS-Visual이 다른 모델에 비해 이미지의 전체적 유사도를 잘 반영함을 검증할 수 있었다. VvsS-Semantic이 가장 낮은 nDCG 점수를 보였고, 시각적 비율을 높임에 따라 더 높은 nDCG 점수를 보임을 확인할 수 있다.

표 4.2로부터 그래프의 의미적 임베딩 모델이 다른 모델들에 비해 높은 nDCG 점수를 확인할 수 있다. VvsS-Visual이 가장 낮은 nDCG 점수를 보였다.

표 4.1와 표 4.2의 결과로부터 본 연구의 모델이 장면 그래프로부터 시각적 특징과 의미적 특징을 어느 수준 성공적으로 분리해서 각각 잘 학습됨을 확인할 수 있다.

Method	Data	nDCG					
		5	10	20	30	40	50
Inter Human	-	-	-	-	-	-	-
ResNet	I	1	1	1	1	1	1
Caption SBERT	Cap(GT)	0.972	0.976	0.980	0.983	0.986	0.988
Gen. Cap. SBERT	Cap(Gen)	0.971	0.974	0.979	0.982	0.985	0.987
Object Count	I+SG	0.971	0.974	0.979	0.982	0.985	0.987
VvsS-Visual	I+SG	**0.979**	**0.981**	**0.984**	**0.986**	**0.988**	**0.990**
VvsS-Semantic	I+SG	0.971	0.975	0.979	0.982	0.985	0.987
VvsS-Half	I+SG	0.973	0.976	0.980	0.983	0.986	0.988
VvsS-Best	I+SG	0.978	0.981	0.984	0.986	0.988	0.990

Table 4.1: 정답 장면 그래프를 이용해 이미지 검색을 진행한 결과로 ResNet-152(시각적 대리 관련도에 사용한 모델) 유사도를 정답 라벨로 간주했을 때의 ndCG 결과

Method	Data	nDCG					
		5	10	20	30	40	50
Inter Human	-	-	-	-	-	-	-
ResNet	I	0.821	0.838	0.859	0.874	0.887	0.898
Caption SBERT	Cap(GT)	1	1	1	1	1	1
Gen. Cap. SBERT	Cap(Gen)	0.823	0.836	0.857	0.872	0.886	0.898
Object Count	I+SG	0.806	0.827	0.850	0.865	0.879	0.895
VvsS-Visual	I+SG	0.805	0.825	0.848	0.864	0.878	0.891
VvsS-Semantic	I+SG	**0.822**	**0.837**	**0.856**	**0.870**	**0.882**	**0.894**
VvsS-Half	I+SG	0.822	0.837	0.856	0.870	0.883	0.895
VvsS-Best	I+SG	0.809	0.829	0.851	0.867	0.880	0.893

Table 4.2: 정답 장면 그래프를 이용해 이미지 검색을 진행한 결과로 정답 캡션 SBERT(의미적 대리 관련도에 사용한 모델)을 정답 라벨로 간주했을 때의 ndCG 결과

3) 인간 동의 점수 결과

대리 관련도에 비교한 nDCG 점수보다 중요한 것은 해당 알고리즘이 인간의 평가와 얼마나 유사한 지를 보이는 인간 동의 점수 결과이다. 이 때 모든 모델들에 대해 최적의 역치값을 정하는 것은 불가능했기에, 알고리즘 간 비교 시에는 쿼리 이미지에 대해 어느 이미지가 가까운지를 무조건 선택하도록 한 결과이다. 표 4.3은 정답 라벨 장면 그래프를 사용했을 때의 인간 동의 점수 결과이다. 시각적 유사도만을 이용하거나 의미적 유사도만을 이용할 때에 비해 이들의 최적 비율값을 사용했을 때 가장 높은 동의 점수를 얻을 수 있었다.

Method	Data	Human agreement score
Inter Human	-	0.728±0.05
ResNet	I	0.494
Caption SBERT	Cap(GT)	0.646
Gen. Cap. SBERT	Cap(Gen)	0.473
Object Count	I+SG(GT)	0.506
VvsS-Visual	I+SG(GT)	0.527
VvsS-Semantic	I+SG(GT)	0.509
VvsS-Half	I+SG(GT)	0.510
VvsS-Best	I+SG(GT)	**0.528**

Table 4.3: 정답 장면 그래프를 사용했을 때의 인간 동의 점수

.

표 4.4은 생성된 장면 그래프를 통해 이미지 검색을 수행한 결과이다. 표 4.3과 비교했을 때 추론 시에 정답 장면 그래프가 아닌 생성된 장면 그래프를 이용했음에도 불구하고 성능이 악화되지 않았다. 반면, 생성된 캡션 SBERT와 실제 캡션 SBERT는 그 성능에 차이가 컸다. 문장의 완성도에 많은 영향을 받는 캡션에 비해 장면 그래프는 생성된 라벨을 이용하더라도 강건하게 대응함을 확인할 수 있었다. 생성된 라벨로 얻은 인간 동의 점수는 정답 캡션 SBERT를 이용한 모델을 제외하고 베이스라인 모델 중에

가장 좋은 성능을 보였다.

Method	Data	Human agreement score
Inter Human	-	0.728±0.05
ResNet	I	0.494
Caption SBERT	Cap(GT)	0.646
Gen. Cap. SBERT	Cap(Gen)	0.473
Object Count	I+SG(Gen)	0.511
VvsS-Visual	I+SG(Gen)	0.501
VvsS-Semantic	I+SG(Gen)	0.523
VvsS-Half	I+SG(Gen)	0.537
VvsS-Best	I+SG(Gen)	**0.537**

Table 4.4: 생성된 장면 그래프를 사용했을 때의 인간 동의 점수

그림 4.7은 시각적 특징과 의미적 특징의 내삽 비율에 따른 인간 동의 점수를 도식화한 결과이다. 베이스라인 모델로는 캡션으로부터 얻은 SBERT feature을 전체 이미지의 ResNet feature와 내삽했다. 실제 정답 캡션 라벨과 ResNet feature을 내삽한 결과가 모든 비율에 대해 가장 좋은 성능을 보였으나, 실 상황에서는 실제 정답 라벨 캡션과 장면 그래프를 가정할 수 없다. 그렇기에 생성된 캡션과 생성된 장면 그래프를 비교했을 때 장면 그래프를 이용했을 때 미미하지만 더 좋은 성능을 보임을 확인할 수 있었다. 사용한 그래프 생성이 이미지 검색이란 작업 목표에 맞게 학습된 것이 아니기에 발전된 장면 그래프 생성 모델을 사용하거나 그래프 노드 생성까지를 종단간 학습으로 학습시킬 경우 결과의 향상이 있을 수 있다.

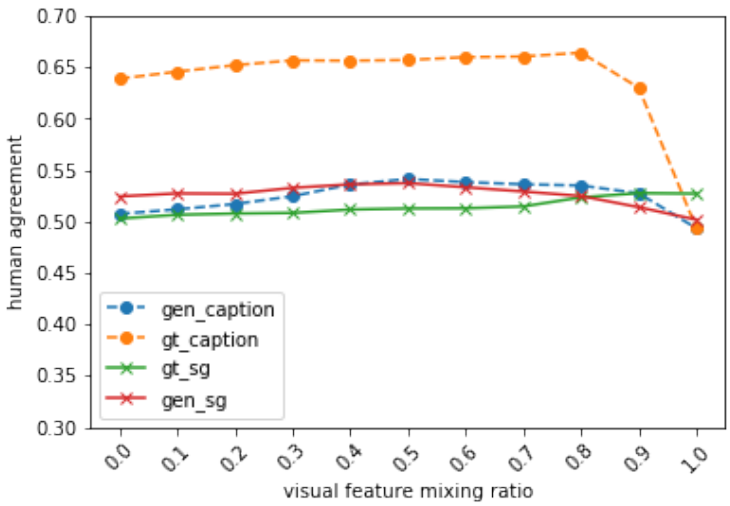


Figure 4.7: 시각적 특징과 의미적 특징의 내삽 비율에 따른 인간 동의 점수

인간 동의 점수를 계산하기 위해 인간이 라벨링한 것과 같이 둘 다 유의미하다 혹은 유의미하지 않다고 판단하는 역치들을 적용했다. 앵커 이미지와 후보 이미지가 유사하다고 판단되는 최소 유사도와 두 개의 이미지가 유사도에 있어서 차이가 있다고 판단되는 최소 마진, 총 2개의 역치가 필요하다. 만약 두 이미지 모두 최소 유사도보다 낮을 경우 유의미하지 않다고 판단하고, 두 이미지 모두 최소 유사도보다 높지만 각 유사도 간에 차이가 최소 마진보다 작을 경우 둘 다 유의미하다고 판단한다. 다양한 역치값 세팅에 대해 실험을 했을 때, 최소 유사도로는 0.4, 최소 마진으로는 0.05가 도출되었다. 다른 역치값 사용에 있어서 인간 동의 점수에 있어서 크게 차이나지 않았다.

또 다른 실험으로 개개인이 이미지 간 유사도를 판단함에 있어서 시각적 유사도와 의미적 유사도에 대한 성향(bias)이 다름을 보여주는 실험을 진행했다. 생성된 장면 그래프의 시각적·의미적 유사도를 λ 비율을 0.0부터 1.0까지 0.1 스케일로 각 비중마다 바꿔가며 인간 동의 점수의 최솟값과 최댓값을 테스트했다. 최대 인간 동의 점수는 0.543이었고, 최소 인간 동의 점수는 0.457이었다. 모든 인간에게 동일한 비중을 적용했을 때 최대 점수가 0.537이었던 것에 비해 개개인마다 다른 최적의 비중을 적용했을 때 0.006 높은 점수를 얻을 수 있었다.

그림 4.8은 각 비중마다 가장 높은 인간 동의 점수를 얻은 사람 수를 나타낸 결과로 개개인이 이미지 검색을 평가함에 있어 다른 패턴을 보임을 확인할 수 있었다.

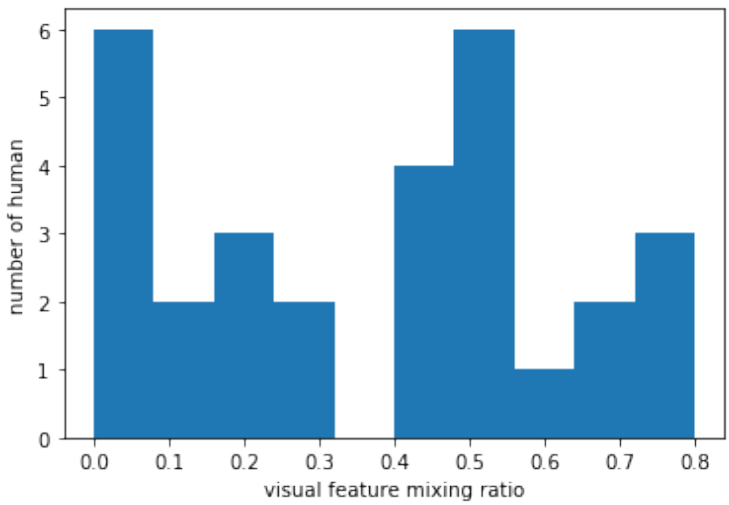


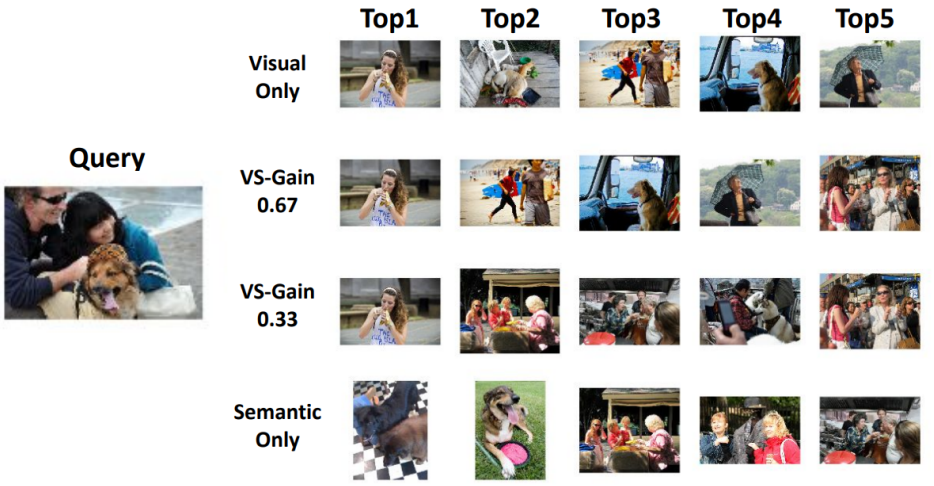
Figure 4.8: 시각적 유사도 비중에 따른 높은 인간 동의 점수를 보인 사람 수

4.3.2 정성적 실험 결과

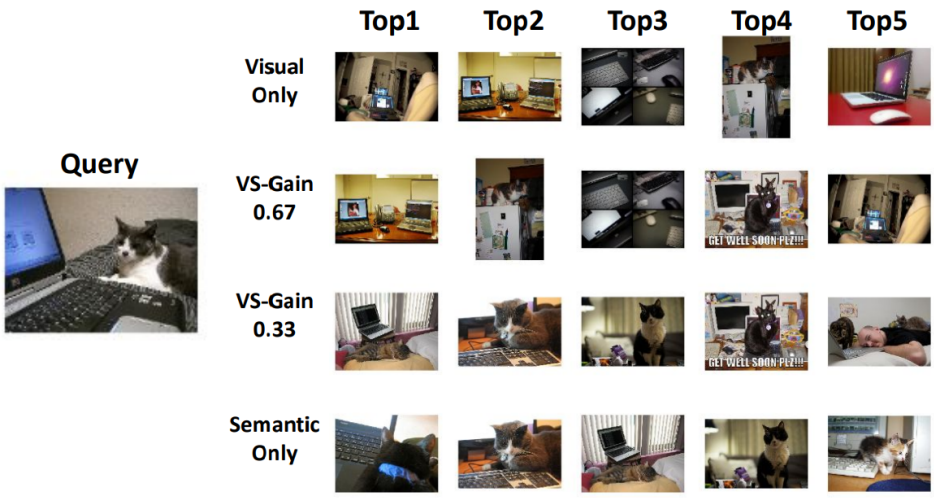
그림 4.9은 쿼리 이미지에 대해 검색된 결과 예시다. 본 연구의 목적은 다른 베이스라인 모델에 비해 좋은 성능을 검증하는 것 역시 있었으나, 최종적인 목적은 VvsS-Net이 시각적·의미적 측면의 비율을 조정해서 검색 가능하다는 가능성을 보이는 것이었다. 각 결과의 첫 번째 행은 오직 시각적 특징만 이용해서 반환한 결과고, 마지막 행은 의미적 특징만 이용해서 반환한 결과이다.

그림 4.9-(a)는 개를 안고 있는 사람들에 대한 이미지를 쿼리 이미지로 검색을 한 결과다. 시각적 특징은 이미지의 패턴에 민감하고 개와 여성이 장면 그래프에서 이어져 있는 노드들이 많기에 높은 중요도를 갖게 된다. 그 결과, 시각적 임베딩만을 중요시하여 반환했을 때는 푸들과 유사한 머리를 가진 여성이 Top-1으로 반환이 되었다. 그러나 의미적 특징은 개와 사람들에 중요도를 주고, 개와 사람들이 존재하는 이미지들을 반환했다. 즉, 패턴에 민감한 시각적 특징을 의미적 특징을 함께 사용함으로써 한계를 극복했다.

그림 4.9-(b)는 VvsS-Net의 가능성을 보여주는 또 다른 예시 중 하나로, 고양이가 노트북 옆에 있는 쿼리이미지로 검색을 한 결과이다. 시각적 유사도에 집중한 결과는 주로 고양이를 무시하고 개체들의 시각적 유사도에 집중해 반환한다. 그러나 의미적 유사도를 반영하면서 노트북 위의 고양이 이미지들을 반환하는 결과를 볼 수 있다. 또한 세 번째 행과 네 번째 행을 비교한 결과, 시각적 유사도를 일부 반영한 결과가 사람이 존재하는 등 의미가 미미하게 다르다고 하더라도 고양이의 시각적 특징을 잘 잡아냄을 확인할 수 있다. 반면 의미적 유사도만 사용했을 경우 고양이의 특징보다는 고양이와 노트북만 존재하는 상황에 집중해 결과를 반환함을 확인할 수 있었다.



(a)



(b)

Figure 4.9: 다른 λ 값에 따른 이미지 검색 결과 예시

제 5 장 결론

5.1 결론

본 논문에서는 시각적·의미적 유사도의 비중을 조정하지 못하는 기존 이미지 검색 연구들의 한계를 벗어나, 시각적·의미적 유사도의 비중을 제어 가능한 이미지 검색의 가능성을 제시하는 VvsS-Net 방법론을 제안했다.

이미지의 ResNet feature, 캡션의 SBERT feature을 대리 관련도로 사용해 개체의 특성을 제외한 장면 그래프의 시각적·의미적 임베딩을 추출하는 그래프 합성곱 신경망을 학습하는 방법을 제안했다. 이후 임베딩의 코사인 유사도를 사용해 이미지 간 유사도를 도출했다.

Visual-Genome 데이터셋과 MS-COCO 데이터셋의 교집합을 사용해 사람이 라벨링한 캡션과 장면 그래프가 모두 갖춰진 이미지 데이터셋을 구축했다. 또한, 이미지 검색 알고리즘 성능의 평가를 위해 이미지 트리플렛에 대해 사람의 유사도 평가 데이터를 수집해서 정량적으로 이미지 검색 성능을 평가할 수 있도록 했다.

각 대리 관련도를 기준으로 그래프의 시각적·의미적 임베딩이 잘 학습됨을 확인했고, 두 정보를 모두 활용할 때 인간 동의 점수에서 높은 성능을 보임을 검증하였다. 또한 장면 그래프의 경우 캡션에 비해 생성된 라벨 역시 강건함을 생성된 캡션과 생성된 장면 그래프의 결과 비교를 통해 검증하였다. 사람이 이미지 간 유사도를 판단할 때 시각적·의미적 측면을 모두 고려하며, 개개인이 각 측면에 느끼는 중요도 역시 다름을 보일 수 있었다.

5.2 향후 연구

본 연구는 기존 모델에 제어 가능한 단위 모듈을 추가하여 시각적·의미적 비율을 제어할 수 있는 가능성을 보이는 것을 목표로 하여, 이미지 검색의 새로운 프레임워크 모델을 처음 제시했다는 의의가 있다. 이 과정에서 특성 노드의 경우 일괄적으로 삭제하는 등 모델을 단순화한 측면이 있다. 그러나 특성 노드의 경우 라벨링한 사람의 성향에 따라 색깔, 질감 같은 단순한 시각적 특징이 아닌 '서 있는', '놓여 있는' 등을 특징으로 가질 수 있다. 그렇기에, 특성 노드에 대해 세심한 처리를 할 경우 모델 성능의 향상의 여지가 있다. 또한 특성 노드를 제거함으로써 시각적·의미적 임베딩의 분리를 의도했으나, 임베딩들을 더 수직적으로 분리하는 기법이 연구될 경우 더 명확한 비율 조절을 기대할 수 있을 것이다.

본 연구에서는 주로 사용자의 의도에 따라 조절함을 가정하고 진행했으나, 이미지의 특성에 따라 그 비율이 달라질 수 있다. 예를 들어, 랜드마크와 같은 쿼리 이미지를 주었을 때는 시각적 특징에 집중한 결과를 반환하거나 복잡한 개체와 관계가 존재하는 쿼리 이미지를 주었을 때는 의미적 특징에 집중한 결과를 판단해서 반환하고자 하는 후속 연구 진행을 기대해볼 수 있을 것이다.

참고 문헌

[1] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[2] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer, 2008.

[3] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016.

[4] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016.

[5] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[6] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. *arXiv preprint arXiv:2201.13182*, 2022.

[7] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[8] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1224–1244, 2017.

[9] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6589–6598, 2017.

[10] Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10718–10726, 2021.

[11] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[12] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[13] Paridhi Maheshwari, Ritwick Chaudhry, and Vishwa Vinay. Scene graph embeddings using relative similarity supervision. *arXiv preprint arXiv:2104.02381*, 2021.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[16] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[17] Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

[18] Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, and Pascal Frossard. Got: an optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems*, 32, 2019.

[19] David Alvarez-Melis and Tommi S Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

[20] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

[21] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.

[22] Lei Zheng, Yang Xiao, and Lingfeng Niu. A brief survey on computational gromov-wasserstein distance. *Procedia Computer Science*, 199:697–702, 2022.

[23] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[27] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[28] Dong Li, Hsin-Ying Lee, Jia-Bin Huang, Shengjin Wang, and Ming-Hsuan Yang. Learning structured semantic embeddings for visual recognition. *arXiv preprint arXiv:1706.01237*, 2017.

[29] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015.

[30] Jia Deng, Alexander C Berg, and Li Fei-Fei. Hierarchical semantic indexing for large scale image retrieval. In *CVPR 2011*, pages 785–792. IEEE, 2011.

[31] Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 638–647. IEEE, 2019.

[32] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[33] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[34] Noa Garcia and George Vogiatzis. Learning non-metric visual similarity for image retrieval. *Image and Vision Computing*, 82:18–25, 2019.

[35] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003.

[36] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010.

[37] Alessandro Bergamo, Lorenzo Torresani, and Andrew W Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*. Citeseer.

[38] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

[39] C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016, 2013.

[40] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011.

[41] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2441–2448, 2014.

[42] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 129–136. IEEE, 2010.

[43] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011.

[44] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[45] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[46] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[47] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[49] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[52] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[53] Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[54] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[55] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[56] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[57] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[58] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[59] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017.

[60] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Abstract

Image-to-Image retrieval with controlling between visual and semantic similarity

SeongEun Lee
Department of Industrial Engineering
The Graduate School
Seoul National University

Image-to-Image retrieval is one of the studies that finds similar images with respect to a query image and has been mainly studied in two approaches: visual similarity and semantic similarity. Since image search is performed in various situations and contexts, there is a limitation to flexibly responding to the user's intention by performing search based on a single criterion.

In this paper, considering both visual and semantic aspects, we propose a methodology that allows users to freely adjust the weight according to the purpose and flexibly search according to the user's intentions. To this end, it is a model that extracts visual and semantic features from a scene graph representing an image through a graph convolutional network, and then interpolates them to adjust the ratio to search. Surrogate relevances of visual feature extracted through pre-trained ResNet-152 model from the image and semantic feature extracted through pre-

trained Sentence-Bert (SBERT) model from human captions are used to train the model.

Through this, image retrieval using learned visual · semantic feature showed high normalized discounted cumulative gain(nDCG) in terms of the surrogate relevance, indicating that each feature was successfully extracted through the graph level. In addition, it was possible to verify that the image search performance was excellent by comparing the quantitative performance with other previous studies in respect to the human agreement score, which shows how similar the algorithm is to human evaluation. In addition, as a result of comparing the visual and semantic features with ResNet and caption SBERT models respectively, it was confirmed that interpolation using features extracted on the same graph level showed better performance. It was also confirmed that this model can successfully perform a search that adjusts the ratio of visual and semantic similarity through the qualitative results of searching images while adjusting the ratio.

Keywords: Image-to-Image Retrieval, Image Similarity, Controllable Image Retrieval, Visual Similarity between Images, Semantic Similarity between Images, Scene Graph, Graph Embeddings, Graph Convolutional Network, Surrogate Relevance, Human Agreement Score, ResNet, Sentence BERT

Student Number: 2019-27796