**텍스트 피드백 기반 이미지 검색 기술 동향**

**시각지능연구실**

**인공지능연구소**

2006-06-311.0

**문서 정보**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **구 분** | **소 속** | **성 명** | **날 짜** | **서 명** |
| **작성자** | 시각지능연구실 | 김가은 | 2022.02.21 |  |
| 시각지능연구실 | 문진영 | 2022.02.21 |  |
| **동료검토자** | 시각지능연구실 | 김형일 | 2022.02.21 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **QA검토자** |  |  |  |  |
|  |  |  |  |
| **승인자** | 시각지능연구실 | 이용주 |  |  |
| **버 전** | 1.0 | | | |
| **발행일** | 2022.02.25 | | | |
| **상 태** |  | | | |

**개정 이력**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **버전** | **개정일자** | **개정 내역** | **작성자** | **확인자** |
| 0.9 | 2022.02.21 | 초안 | 김가은, 문진영 | 이용주 |
|  |  |  |  |  |
|  |  |  |  |  |

목 차

[1. 개요 4](#_Toc96330717)

[2. 텍스트 피드백 기반 이미지 검색 문제 정의 4](#_Toc96330718)

[3. TIRG 5](#_Toc96330719)

[4. MAAF 8](#_Toc96330720)

# 개요

* 1. 목적

본 문서는 컴퓨터 비전 분야에서 제안된 텍스트 피드백 기반 이미지 검색 기술에 대한 최신 연구 동향과 각 연구에서 제시하는 이미지 검색 모델에 대한 설명을 기술한다.

* 1. 범위

본 문서는 텍스트 피드백 기반 이미지 검색 문제에 대한 정의와 최신 연구에서 제시한 세 개의 검색 모델에 대한 설명을 포함한다.

[1] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays. Composing text and image for image retrieval – an empirical odyssey. In IEEE Conference on Computer Vision and Pattern Recognition, 2019. 🡪 TIRG

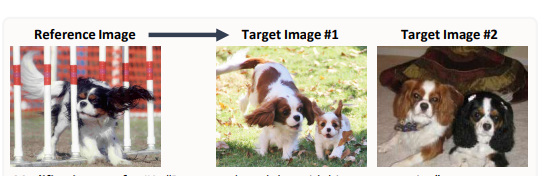
[2] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakye. Modality-agnostic attention fusion for visual search with text feedback. ArXiv, abs/2007.00145, 2020. 🡪 MAAF

[3] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In IEEE International Conference on Computer Vision, 2021. 🡪 CIRPLANT

# 텍스트 피드백 기반 이미지 검색 문제 정의

우선 이미지 검색 문제는 사용자가 이미지를 찾고자 할 때, 공식화된 검색 query를 이용하여 이미지를 검색한다. 이 때, 검색 쿼리로 공식화하는 방법에는 텍스트 문자열, 유사한 이미지, 스케치 또는 이들의 조합 등을 사용하여 나타낸다. 본 문서에서 다루는 텍스트 피드백 기반 이미지 검색 문제는 쿼리가 입력 이미지와 이미지에 대한 일부 원하는 수정을 설명하는 텍스트 문자열로 공식화되는 경우를 사용한다.

텍스트 피드백 기반 이미지 검색은 딥러닝 분야 중 이미지 처리와 텍스트 마이닝의 혼합모형으로, 미지와 언어 쌍으로 제공된 사용자 쿼리와 가장 일치하는 대규모 말뭉치에서 이미지를 찾는 작업을 목표로 한다.



Modification text for #1: “Be a same breed dog with his puppy running”

Modification text for #2: “Two dogs of the same breed on the floor”

그림 1. 텍스트 기반 이미지 검색의 예시

그림 1을 보면 알 수 있듯, 합성 이미지 검색의 입력(input query)은 이미지와 짧은 텍스트 설명으로 구성되어 있다. 이 query를 학습한 후 텍스트로 그 쿼리에 맞는 이미지를 찾는 합성 이미지 검색 작업을 요구한다.

* 1. 기존 다른 방법과의 차이점

본 문서에서 다루는 문제는 단일 모딜리티가 사용되는 기존의 콘텐츠 기반 또는 텍스트 기반 이미지 검색과는 다르다고 할 수 있다. 단일 모딜리티가 사용되는 것의 예는 비슷한 이미지를 input으로 두어 원하는 이미지를 검색하는 것이 있다. 반면, 텍스트 기반 이미지 검색은 사용자의 의도를 명시하기 위해 시각적 및 텍스트 모달리티를 모두 포함한다. 이러한 멀티모달리티의 장점은 어떤 개념과 속성은 시각적으로 더 간결하게 묘사되고, 그 외에는 언어를 통해 설명된다. 두 모달리티 상호 reference함으로써 기준 영상은 씬(scene)의 일반적인 요지를 캡쳐할 수 있으며 텍스트는 더 자세한 내용을 지정할 수 있다.

# TIRG

관련 논문

[1] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays. Composing text and image for image retrieval – an empirical odyssey. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

TIRG는 텍스트 피드백 기반 이미지 검색 문제를 해결할 첫 번째 기술로, 새로 제안된 모델이다. 이미지와 텍스트라는 두 가지 입력양식을 가지고 있을 때, 이를 쿼리라고 두고, 이 쿼리에 대한 의미있는 교차 모달 피쳐 합성(cross-modal feature composition)을 학습을 하여 target 이미지를 찾음으로써 이미지 검색 문제를 해결한다. 이는 후에 질의 분류, 작문 학습 등과 같은 문제에서도 성공적으로 사용된다.

* 1. 선행 기술

TIRG 이전에는 텍스트 피드백 기반 이미지 검색 문제를 해결하는 선행 기술이 존재하지 않는다. 하지만 유사한 다른 분야의 선행 기술들을 비교해볼 수 있다. 공정한 비교를 위해, 다 같은 pipeline을 사용하여 모든 방법을 교육하지만 구성 모듈만 다르게 하여 진행할 것이다. 첫 번 째는 단일 모달리티를 사용하는 text only, Image only가 있다. 두 번째는 두 개의 MLP 층과 ReLU를 사용하여 text와 image의 계산을 연결하는 방법이 있다. 세 번째 선행기술은 Relationship은 관계적 재구성을 포착하는 방법이다. 이는 CNN을 사용하여 이미지로부터 2D맵 추출하고, 텍스트 feature와 이미지 feature세트를 MLP를 통해 전달하고 평균화 되어 단일 피쳐를 생성하는 방법이다.

* 1. 네트워크 구조

이 작업의 목표는 수정된 이미지와 target 이미지의 feature의 임베딩을 더 가깝게 하는 동시에 유사하지 않은 이미지의 feature을 분리하는 것이다. 분리할 때에는 분류손실을 사용한다. 임베딩을 가깝게 하는 과정을 살펴보자.

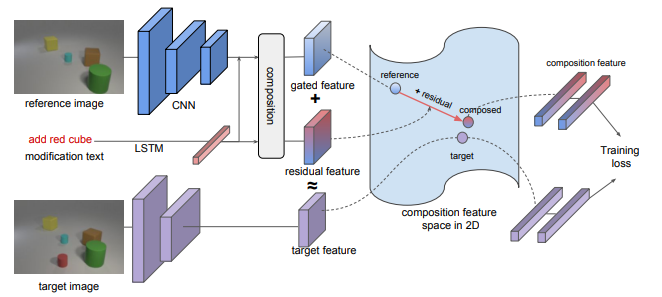


그림 2.

다음 그림은

먼저 reference 이미지는 ResNet-17 CNN을 사용하여 인코딩을 진행(하고, reference 텍스트는LSTM을 사용하여 인코딩()한다. 마지막으로 인코딩된 두 기능을 결합하여를 계산한다. 는 텍스트 이미지 잔차라고 부르는 TIRG접근법을 사용하여 이미지와 텍스트 기능을 결합해주는 함수이다. 그림2를 보면 구성 모듈에는 잔류연결과 게이트 연결의 두 가지 연결 유형이 있는 것을 볼 수 있다. 따라서, 게이팅연결과 잔차연결을 계산한 후 덧셈을 활용해 두 형상벡터를 결합한다. 식은 다음과 같다.



덧셈으로 다른 feature공간을 만들기 보다, 텍스트 피쳐를 기반으로 이미지 기능을 수정하게끔 한다. 이는 taget 피쳐와 비슷해지고, 2D공간에 매핑되어 loss값의 input이 된다.

* 1. triplet loss

위에서 유사하지 않은 이미지의 feature을 분리할 때에는 분류손실을 사용한다고 하였다.

를 마지막 층의 최종 이미지-텍스트 쿼리 표현이고, 를 그 쿼리의 target 이미지 표현이라고 하자. 만약 쿼리 수가 B개일 때, soft triplet loss를 다음과 같이 정의할 수 있다.



* 1. 선행기술 대비 장점 및 보안

Fashion 200k 데이터셋을 사용하였을 때, TIRG가 기존의 접근 방식들을 능가한다는 것을 알 수 있다. 이는 이미지와 텍스트를 같은 공간에 삽입한 선행기술들과는 다르게 이미지 공간을 수정하기 위해 텍스트를 사용하였기 때문에 좋은 결과값이 나온 것이라 생각된다. 이러한 장점을 더 살리기 위해 TIRG에서는 기존 벤치마크에는 없었던 복잡한 테스트 수정 사항을 포함하는 새 데이터 세트 CSS를 만들었다. 이 데이터셋을 사용하니 feature 수정의 효과가 더욱 부각되었다.

잔류연결과 게이트 연결의 중요성을 보여주는 실험을 진행하였고, 그 결과 잔류 feature나 게이트 feature를 제거하면 성능이 저하되는 것을 볼 수 있다. 따라서 이미지 잔차와 게이트는 TIRG모델에 있어서 중요 성능이라고 할 수 있다.

# MAAF

MAAF모델은 세분화된 시각적 기능을 기반으로 catalog검색을 제공한다. 따라서 이 모델은 이미지 및 텍스트 기능을 결합하고 Fashion IQ 및 CSS 데이터셋의 텍스트 데이터셋을 수정하여 단일 단어 수정만으로 수행할 것이며, 본 장에서 TIRG 모델의 수정이 없이 진행할 것이다. 한 멀티모달 이미지 검색 작업에서 모델을 평가하기 위해 사용하는 풍부한 언어입력으로 새로운 설정을 제공하는 두 가지 새로운 벤치마크(Birds-to-Words, Spot-the-Diff)를 소개할 것이다

* 1. 선행 기술의 문제

선행 기술은 TIRG은 이미지와 텍스트 모두를 포함하는 보다 복잡한 다중 모드 쿼리를 허용하였다. 이러한 시스템은 이미지 검색 target을 표현력 있게 전달한다. 그래서 TIRG에서는 이미지를 수정하는 방법을 설명하는 텍스트와 함께 쿼리가 이미지에 의해 지정되는 설정을 연구하였다. 하지만 패션 catalog검색과 같은 중요 응용프로그램이 포함되었을 때, 자연 언어의 유연성이 적합하다. 따라서 TIRG도 좋은 방안이지만, 패션 catalog검색의 경우 텍스트와 이미지 특징을 잔여 동기 메커니즘과 결합하면 더 좋은 결과를 낼 수 있다.

* 1. 구조

MAAF모델의 구조를 간단히 설명하자면, 텍스트 수정 이미지 검색 작업을 해결하기 위해 각 모달리티(텍스트와 이미지)에 대해 토큰벡터를 추출한 다음, 토큰을 단일 시퀀스로 쌓아 attention모델로 처리하는 것이다. attention모델의 출력 값을 가장 가까운 이웃 검색(nearest-neighbor search)에 적합한 단일 벡터로 결합해야 한다.

모델을 조금 더 깊이 설명하자면, 텍스트에 의해 수정된 쿼리 이미지와 카탈로그 이미지에 대한 공통 임베딩 공간을 학습하여 주어진 이미지+텍스트 쿼리에 대해 가장 밀접하게 일치하는 카탈로그 항목의 순위를 매긴다. 이때 순위는 해당 쿼리 임베딩과 카탈로그 이미지의 임베딩 사이의 유사성을 계산하여 매겨진다.

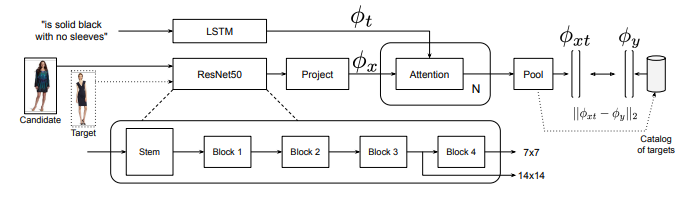


그림 3.

이미지는 ResNet50으로 전달되고, 텍스트는 토큰화된 후 LSTM으로 전달된다. 각 모달리티에서 토큰 시퀀스를 추출하고 그 결과를 트랜스포머 인코더에서와 같이 dot-product attention을 사용하고 fully-connected layers를 사용해 이들을 융합한다.

self-attention을 통해 이미지-이미지, 텍스트-텍스트, 또 cross attention을 통해 이미지-텍스트 상호작용을 허용한다. Self-attention은 각 토큰이 다른 토큰을 볼 수 있도록 허용하므로, 이미지 토큰은 다른 이미지 토큰 또는 텍스트 토큰에 의해 가장 강력하게 수정될 수 있다. Cross attention은 이미지 토큰을 하나의 시퀀스로 처리하고 텍스트 토큰을 다른 시퀀스로 처리하는 방법이다. 이러한 방법은 CSS작업에서 훨씬 더 나은 성능을 얻을 수 있다.

* 1. Loss

삼중항 loss를 사용하면 recall@k metric에 의해 나쁜 성능을 보인다. 따라서, i번째 쿼리의 로스는 다음과 같이 정의된다.



* 1. 벤치마크

Birds-to-Words는 새의 이미지와 이미지들 사이의 차이를 설명하기 위해 쓴 텍스트로 구성되어 있다. 각각은 평균 32.1개의 토큰에서 한 쌍의 새들 사이의 차이를 설명한다. Birds-to-Words는 예제 수는 적지만 기존 연구하는 다른 데이터 세트보다 각 예에서 더 풍부한 텍스트 설명을 제공한다. 상대 캡션이 주어진 경우 설명된 두 번째 이미지를 검색하는 작업에 맞게 데이터 세트를 조정하였다. 기준 모델은 공유 구성 요소에 대해 동일한 매개 변수로 훈련되며, 검증 세트 성능을 최적화하기 위해 교육 반복 횟수를 선택하였다.

Spot-the-Diff 감시 영상의 스냅샷과 인간이 한 쌍의 차이를 설명하기 위해 작성한 단일 문장 캡션으로 구성된다. 이 데이터 세트는 원래 비교상의 캡션 생성을 연구하기 위해 설계되었지만, 동일한 소스의 이미지가 매우 유사하기 때문에 maaf 작업에 어려운 설정을 제공한다.

* 1. 선행 기술 대비 장점

선행 기술인 TIRG보다 더 나은 성능을 보이는데, 이는 위에서 언급했듯 패션 catalog검색과 같은 중요 응용프로그램이 포함되었을 때, 텍스트와 이미지 특징을 잔여 동기 메커니즘과 결합으로 성능을 더 높이기 때문이다. 따라서, 모델의 성능 향상 원인을 더 잘 파악하기 위해 색상(빨간색, 파란색, 녹색 등)과 텍스처(스트라이프, 솔리드, 격자 등...)와 같은 여러 유형의 키워드를 포함하는 패션 IQ 검증 세트의 하위 집합에 대해 평가를 진행하였다. 격자무늬 블루와 블랙과 같은 캡션은 여러 한정자 클래스(이 경우 격자무늬 텍스처와 파란색과 검은색)에 대한 키워드를 포함하므로 이러한 한정자 클래스는 상호 배타적이지 않다. 모든 하위 집합에 대해, MAAF 모델은 (Recall@10 + Recall@50)/2에 비해 약 16% 정도 TIRG를 능가한다.

또한 MAAF모델이 validation set에서 다른 모든 단일 모델보다 성능이 우수하다는 것을 발견했다. [27]의 모델은 TIRG[47]와 매우 유사하고 이 추가 비공개 데이터 없이 훈련된 모델에 대한 결과를 사용할 수 없기 때문에, 우리는 이 텍스트 말뭉치가 TIRG에 비해 향상된 성능에 중요하다고 추측한다. 또한 TIRG와 직접 비교해보니 우리 모델이 더 나은 성능을 보인다. 전체 결과는 표 2와 표 3을 참조하라. 여기에는 주의 메커니즘에 대한 설계 선택의 세부 사항이 포함된다

# CIRPLANT

MAAF모델은 세분화된