**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 | 2022.01 |  |
| 시각지능연구실 | 문진영 | 2022.01 |  |
| **동료검토자** | 시각지능연구실 | 김형일 | 2022.01 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **QA검토자** |  |  |  |  |
|  |  |  |  |
| **승인자** | 시각지능연구실 | 이용주 |  |  |
| **버 전** | 1.0 | | | |
| **발행일** | 2022.02.25 | | | |
| **상 태** |  | | | |

**개정 이력**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **버전** | **개정일자** | **개정 내역** | **작성자** | **확인자** |
| 0.9 | 2022.02.07 | 초안 | 김가은, 문진영 | 이용주 |
|  |  |  |  |  |
|  |  |  |  |  |

목 차

[**1. 개요 4**](#_Toc50993987)

1.1 목적………………………………………………………………………………………………………………………………………….4

1.2 범위………………………………………………………………………………………………………………………………………….4

[**2.** **텍스트 피드백 기반 이미지 검색** **문제 정의** 4.](#_Toc50993988)

2.1 기존의 다른 문제와의 차이점……………………………………………………………………………………………………….5

[**3. TIRG 10**](#_Toc50993989)

[**5. MAAF 11**](#_Toc50993991)

[**5. CIRPLANT 11**](#_Toc50993991)

[**6. Reference 12**](#_Toc50993992)

# 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.

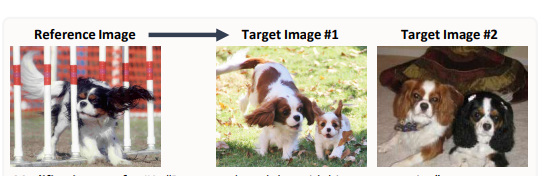
* 1. 범위

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

# 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. 기존의 다른 문제와의 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

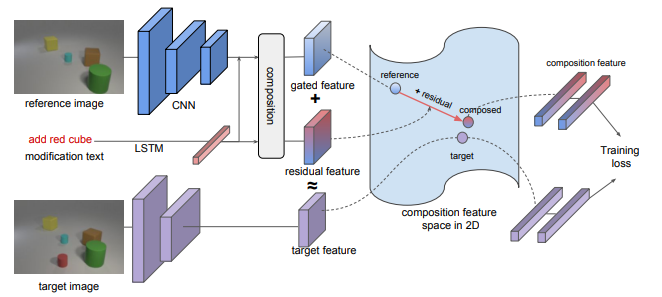
Tirg는 텍스트 피드백 기반 이미지 검색 문제를 해결할 첫 번째 기술로, 새로 제안된 모델이다. 이미지와 텍스트라는 두 가지 입력양식을 가지고 있을 때, 이를 쿼리라고 두고, 이 쿼리에 대한 의미있는 교차 모달 피쳐 합성(cross-modal feature composition)을 학습을 하여 taget 이미지를 찾음으로써 이미지 검색 문제를 해결한다. 이는 후에 질의 분류, 작문 학습 등과 같은 문제에서도 성공적으로 사용된다.

* 1. 선행기술

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. 구조

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



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 모델에 있어서 중요 성능이다.

# MAFF

MAFF모델은 이미지 및 텍스트 기능을 결합하고 Fashion IQ 및 CSS 데이터셋의 텍스트 데이터셋을 수정하여 두 가지 시각적 검색에 대한 기존 접근방식을 능가하며, 단일 단어 수정만으로 수행한다. 수정 없는 접근 방식이 더 우월하다는 것을 보여줄 것이다.

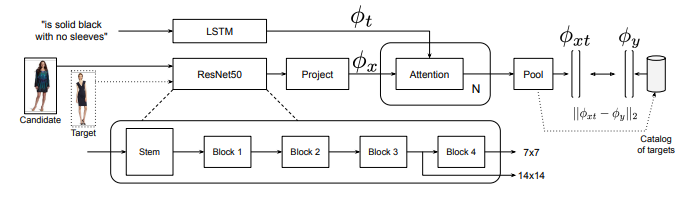
Maff는 각 모달리티에 대해 입력의 일부를 나타내는 “토큰”벡터를 추출한 다음 이러한 토큰을 단일 시퀀스로 쌓아 주의 모델로 처리하는 것이다. 이러한 각 단계에 대한 접근 방식을 비교하고 텍스트 입력을 위한 간단한 LSTM, 이미지 입력을 위한 여러 척도의 CNN, 각 토큰 클래스 내의 평균 풀링이 가장 강력한 결과를 달성함을 보일 것이다. 멀티모달 이미지 검색 작업에서 모델을 평가하기 위한 두 가지 새로운 벤치마크도 사용한다.

* 1. 선행 기술의 문제

선행 기술은 tirg보다

* 1. 구조

텍스트에 의해 수정된 쿼리 이미지와 카탈로그 이미지에 대한 공통 임베딩 공간을 학습하여 주어진 이미지+텍스트 쿼리에 대해 가장 밀접하게 일치하는 카탈로그 항목의 순위를 매겨 해당 쿼리 임베딩과 카탈로그 이미지의 임베딩 사이의 유사성을 계산하여 매기는 것이다.



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. 선행 기술 대비 장점

선행 기술인 tirg보다 더 나은 성능을 보이는데, 이는