

## Image Captioning through Image ‘Transformer’

Sen He\*!, Wentong Liao\*”?!, Hamed R. Tavakoli\*®®, Michael Yang\*  
Bodo Rosenhahn?, Nicolas Pugeault®

\* CVSSP, 영국 서리 대학교

“라이프니츠 대학은 독일 하노버 대학이다.

\* 노키아 테크놀로지스, 핀란드

네덜란드 트렌트 대학은 2016년 1월에 1년간 1억 유로(약 1억 2천 600만 원)를

” 컴퓨팅 과학 학교, 글래스고 대학

senhe/52@gmail.com

초록. 이미지의 자동 자막화는 이미지 분석과 텍스트 생성의 과제를 결합한 과제이다. 자막화의 중요한 특징 중 하나는 주의 집중 개념, 즉 어떤 것을 어떻게 묘사할 것인지, 어떠한 순서로 묘사할 것인지 결정하는 것이다. 텍스트 분석과 번역의 성공에서 영감을 얻어 이전의 작업들은 이미지 캡션을 위한 트랜스포머 아키텍처를 제안했다. 그러나 이미지(일반적으로 객체 검출 모델에서 검출된 영역)와 문장(각 단일 단어)에서 시멘틱 단위 사이의 구조는 다르다. 변환기의 내부 아키텍처를 이미지에 적응시키기 위한 제한된 작업이 이루어져 왔다. 이 작업에서, 우리는 이미지 영역들 간의 상대적인 공간 관계에 의해 동기화된 수정된 인코딩 변환기와 암시적 디코딩 변환기를 구성하는 이미지 변환기를 소개한다. 우리의 디자인은 원래의 변환기 레이어의 내부 아치 구조를 넓혀 이미지의 구조에 적응한다. 입력으로서의 지역만을 갖는 모델은 MSCOCO 오프라인 및 온라인 테스트 벤치마크 모두에 새로운 최첨단 성능을 달성한다. 코드는 <https://github.com/wtliao/ImageTransformer>에서 사용할 수 있다.

### 1 Introduction

이미지 캡션은 단어로 이미지의 내용을 기술하는 작업으로 최근 수년간 언어와 이미지 처리 모두를 위한 딥 러닝 모델의 성공으로 인해 AI 시스템에 의한 자동 이미지 캡션의 문제가 많은 관심을 받아왔다. 문헌에서 대부분의 이미지 캡션화 접근법은 병진 접근법을 기반으로 하고, 시각 인코더와 언어 디코더가 있다. 자동 번역의 한 가지 과제는 단어 단위로 할 수 없다는 것인데, 다른 단어는 그 의미를 영향을 미치고 따라서 단어의 번역이 가능하지 않다는 것인데, 시스템이 이미지에서 설명해야 하는 것을 결정해야 할 이미지에서 텍스트로 번역하는 경우에는 더욱더 그렇다. 이 문제에 대한 일반적인 해결책은 주의 메커니즘에 기초한다. 예를 들어, 이전 이미지 캡션 모델은 해결하려고 한다.

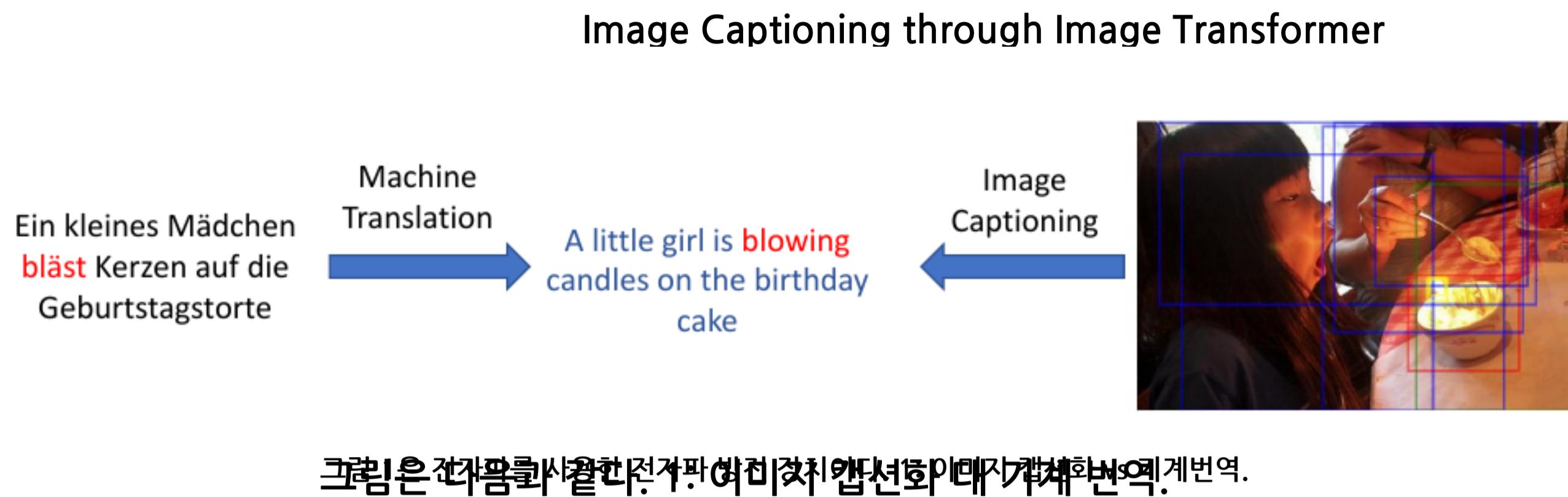
\* Equal contribution

영상 [1-4]에서 어디서 찾아야 하는지(이제 부분적으로 Faster-RCNN 객체 검출 모델 [5]로 해결됨) 인코딩 단계에서 주의 메커니즘을 갖는 순환 신경망을 복호화 단계에서 이용하여 캡션을 생성한다. 하지만 이미지에서 무엇을 설명할지를 결정하기 위한 것 이상으로, 최근의 이미지 캡션-인 모델은 이미지의 영역이 서로 관련되는 방법을 학습하도록 주의를 사용하여 이미지에 그들의 컨텍스트를 효과적으로 인코딩할 것을 제안한다. 그래프 컨볼루션 신경망[6]이 이미지 내의 영역을 관련시키기 위해 처음으로 도입되었지만, 그들 접근법[7-10]은 일반적으로 보조 모델(예를 들어, 보조 모델)을 필요로 한다. 비주얼 관계 검출 및/또는 속성 검출 모델(들)을 사용하여, 애초에 이미지 내 비주얼 장면 그래프를 구축한다. 대조적으로, 자연어 처리 분야에서는 문장 내의 임베디드 단어들을 연관시키기 위해 변환기 아키텍처[11]가 개발되었으며, 그러한 관계들을 명시적으로 감지하는 보조 모델들 없이 단일 층단으로 트레이닝될 수 있다. 최근의 이미지 캡션 모델 [12-14]은 최첨단 성능을 달성하는 내적 곱 어텐션을 통해 이미지 내 정보적인 영역을 암시적으로 연관시키기 위해 변환 아키텍처를 채택했다.

그러나 변환기 아키텍처는 텍스트의 기계 번역을 위해 설계되었다. 문자에서, 한 단어는 다른 단어의 왼쪽이나 오른쪽에 있는데, 거리가 다르다. 대조적으로, 이미지들은 2차원(실제, 3차원 장면들을 나타내기)이므로, 영역이 다른 영역의 좌측 또는 우측에 있을 수 있을 뿐만 아니라, 다른 영역을 포함하거나 또는 포함될 수 있다. 영상의 의미 단위 간의 상대적 공간적 관계는 문장에서의 그것보다 자유도가 크다. 더욱이, 기계 번역의 디코딩 단계에서는, 단어가 통상적으로 다른 언어의 다른 단어로 번역(일대일 디코딩)하는 반면, 이미지 영역에 대해서는 그 맥락, 그 속성 및/또는 다른 영역과의 관계를 기술할 수 있다(일대다 디코딩). 이전 트랜스포머 기반 이미지 캡션 모델[12-14]의 한 가지 제한은 트랜스포머 계층이 단일(멀티헤드) 도트-프로덕트 어텐션 모듈을 포함한다는 것인데, 이 모듈은 기계 번역을 위해 설계된 트랜스포머의 내부 아키텍처를 채택한다. 본 논문에서는 이미지 캡션을 위한 이미지 변환기를 소개하는데, 여기서 각 변환계층은 다수의 서브 변환기를 구현하여 이미지 영역 간 공간 관계를 부호화하고, 이미지 영역 내 다양하게 정보를 복호화한다.

우리 방식과 이전 변환기 기반 모델 [12, 14, 13] 간의 차이점은 우리 방식이 변환기 모듈을 넓혀주는 트랜스-포머 계층의 내부 아키텍처에 초점을 맞춘다는 것이다. Yao 등은 이러한 결과를 보았다. [10]는 그들의 모델의 인코딩 부분에서 계층적 개념을 사용했으며, 그들의 방법은 글로벌 트리 계층화인 반면, 우리의 모델은 각 질의 영역에 대한 국부적 공간 관계에 초점을 맞춘다. 더욱이, 우리의 모델은 보조 모델들(즉, 시각적 관계 검출 및 인스턴스 세그먼트화에 대해)가 시각적 에네 그래프를 구축하는 것을 요구하지 않는다. 우리의 인코딩 방법은 보조적 관계 및 속성 검출기 없이 암시적으로 조합하는 변환층을 사용하는 시각적 의미 그래프 및 공간 그래프의 조합으로 볼 수 있다.

본 논문의 기여는 다음과 같이 요약될 수 있다.



— 우리는 이미지 영역의 복잡한 자연 구조에 적합한 수정된 주의 모듈을 갖는 이미지 캡션 태스크에 적응된 변환기 층에 대한 새로운 내부 아키텍처를 제안한다.

—we 보고된 제안된 아키텍처를 검증하기 위해 철저한 실험과 절제 연구가 작업에서 수행되었으며, 지역 특징만을 입력으로 사용하여 MSCOCO 이미지 캡션의 오프라인 및 온라인 테스트 데이터세트에서 최첨단 성능을 달성했다.

## 2 Related Work

우리는 현재 주의 기반 이미지 캡션 모델을 단일 단계 주의 모델, 2단계 주의 모델, 시각 장면 그래프 기반 모델, 트랜스포머 기반 모델로 특성화한다. 이 절에서는 이들에 대해 하나씩 검토해 볼 것이다.

## 2.1 Single-Stage Attention Based Image Captioning

단계별 주의 기반 이미지 캡션 모델은 디코딩 단계에서 주의를 적용하는 모델이며, 이 모델은 해당 단어를 생성할 때 이미지 내의 가장 정보적인 영역을 디코더가 주목하는 것[15]이다.

대규모 주석 데이터세트[16,17]의 가용성은 이미지 캡션을 위한 심층 모델의 훈련을 가능하게 하였다. Vinyals et al. [18]은 이미지 캡션을 위한 최초의 딥 모델을 제안했다. 그들의 모델은 이미지를 인코딩하기 위해 ImageNet[16]에 대해 미리 훈련된 CNN을 사용하고, 그 다음 LSTM[19] 기반 언어 모델을 사용하여 이미지 특징을 단어의 시퀀스로 디코딩한다. Xu 등은 이와 같은 결과를 보고한 바 있다. [1]은 언어 모델의 숨겨진 상태 및 이전 생성된 단어를 기반으로 각 단어의 생성 과정에서 이미지 캡션에 주의 메커니즘을 도입하였다. 이들의 주의 모듈은 인코딩된 특징맵 내 각 수용필드를 가중치로 하기 위해 매트릭스를 생성한 다음, 가중치가 부여된 특징맵과 이전 생성된 단어를 언어 모델에 공급하여 다음 단어를 생성한다. 인코딩된 특징 맵의 수용 영역만을 다루는 대신, Chen 등은 수용 영역을 다루고 있다. [2]는 특징 채널 주의 모듈을 추가했는데, 그것은 그들의 채널 주의 모듈이다.

각 워드의 생성 시 각각의 특징 채널을 다시 가중하기(re-weight)한다. 문장에 들어 있는 모든 단어가 이미지에 대응하는 단어가 아니기 때문에 Lu 등은 이를 고려하였다. [20]은 적응적 주의집중 접근법을 제안했는데, 그들의 모델은 시각적 정보에 의존할 때와 장소가 적응적으로 결정되는 시각적 감시원이 있다.

단일 단계 주의 모델은 계산적으로 효율적이지만 원본 이미지에서 정보 지역의 속도 위치 설정이 부족하다.

## 2.2 Two-Stages Attention Based Image Captioning

Two 스테이지 어텐션 모델은 하향-업 어텐션과 상향-다운 어텐션으로 구성되며, 하향-업 어텐션은 먼저 객체 검출 모델을 사용하여 이미지에서 여러 정보 영역을 검출한 다음 단어 생성 시 가장 관련 있는 검출된 영역에 주의한다.

단일 단계 주의 모델과 같이 이미지의 정보 영역으로서 거친 수용 영역에 의존하는 대신에, Anderson et al. 등은 거친 수용 영역을 사용한다. [3]은 Visual Genome 데이터 세트 [21]에서 탐지 모델을 훈련시킨다. 학습된 검출 모델은 이미지에서 10 내지 100개의 정보적인 영역을 검출할 수 있다. 그런 다음 이들은 디코더로서 2-층 LSTM 네트워크를 사용하는데, 여기서 첫 번째 층은 임베디드된 단어 벡터와 검출된 영역의 평균 특징을 기반으로 상태 벡터를 생성하고 두 번째 층은 이전 층으로부터 상태 벡터를 사용하여 검출된 영역 각각에 대한 가중치를 생성한다. 검출된 영역들의 가중합은 다음 단어를 예측하기 위한 컨텍스트 벡터로 사용된다. Lu 등은 2003년 9월에 발표한 연구 결과를 참고하였다. [4]는 비슷한 네트워크를 개발했지만 비주얼 지노미보다 더 작은 데이터 세트인 MSCOCO [22]에 대해 훈련된 검출 모델을 가지고 있었고, 따라서 더 적은 정보를 제공하는 영역이 검출되었다.

2단계 주의 집중 기반 이미지 캡션 모델의 성능은 단일 단계 주의 집중 기반 모델에 비해 많이 향상되었다. 그러나 검출된 각 영역은 다른 영역과의 관계를 갖지 못한 채 다른 영역과 격리되어 있다.

## 2.3. Visual Scene Graph Based Image Captioning

시각적 장면 그래프 기반 이미지 캡션 모델은 검출된 정보적인 영역을 관련시키기 위해 그래프 컨볼루션 신경망을 주입하여 2단계 주의 모델을 확장하고, 따라서 디코더에 입력하기 전에 그들의 특징을 정제한다.

Yao 등은 이와 같은 연구 결과를 발표하였다. [7]는 시맨틱 장면 그래프와 공간 장면 그래프로 구성된 모델을 개발하였다. 의미론적 장면 그래프에서, 각각의 영역은 다른 의미론적으로 관련되는 영역과 연결되고, 그 관계는 통상적으로 합합 박스 중에서 시각적 관계 검출기에 의해 결정된다. 공간 장면 그래프에서 두 영역의 관계는 그들의 상대적인 위치에 의해 정의된다. 그런 다음, 그래프 신경망[6]을 통해 장면 그래프 내의 각 노드의 특징이 관련 노드와 함께 구체화된다. 양 등은 그중에서 가장 많은 비율을 차지하였다. [8]은 자동 인코더를 사용하는데, 이들은 먼저 SPICE [23] 평가 메트릭을 기반으로 문장의 그래프 구조를 인코딩하여 사전을 학습한 다음, 학습된 사전을 사용하여 의미적 장면 그래프를 인코딩한다. 이전 두 개의 작업은 시너그라프에서 의미 관계를 에지로서 다루는 반면 Guo 등은 의미 관계를 에지로서 다루고 있다. [9]에서는 장면의 노드로 다루고 있다.

## Image Captioning through Image Transformer on

그래프를 그려보세요. 또한, 이들의 디코더는 영역의 다른 측면에 초점을 맞춘다. Yao et al. [10]은 트리 계층 및 인스턴스 레벨 피처를 장면 그래프에 추가로 도입한다.

정보성 영역을 관련지으려는 그래프 신경망을 도입하면, 2단계 주의 모델에 비해 이미지 캡션 모델에 대한 상당한 성능 향상을 얻을 수 있다. 단, 초기에는 장면 그래프를 검출하고 구축하기 위한 보조 모델이 필요하다. 또한 그 모델들은 통상적으로 두 개의 평행한 스트림을 갖고 있는데, 하나는 의미론적 장면 그래프를 담당하고 다른 하나는 공간적 장면 그래프를 담당하며, 이는 계산 비효율적이다.

### "Transformer Based Image Captioning"

트랜스포머 기반 이미지 캡션 모델은 정보적인 영역을 암시적으로 관련시키기 위해 내적 곱 집합 주의 메커니즘을 사용한다.

원래 변환 모형[11]이 도입된 이후, 문장의 구조 또는 자연적 특징[24-26]에 기초한 기계 번역을 위해 더 발전된 아키텍처가 제안되었다. 이미지 캡션에서 AoANet[12]는 다중 헤드 주의의 상단에 게이트된 선형 계층[27]을 추가하여 원래 내부 트랜스포머 계층 아키텍처를 사용한다. 객체 관계 네트워크(14)는 상대적 공간 주의를 점 곱 주의에 주입한다. Herdade 등에 의해 기술된 또 다른 흥미로운 결과이다. [14]은 간단한 포지션 인코딩(원래의 변환기에서 제안된 바와 같이)이 이미지 캡션을 개선하지 못했다는 것이다. 연관된 변환기 모델 [13]은 게이트된 양방향 컨트롤러를 통해 융합되는 이미지에서 시각적 및 의미 정보를 인코딩하고 정교화하기 위한 이중 병렬 변환기를 특징으로 한다.

장면 그래프 기반의 이미지 캡션 모델에 비해, 변환기 기반의 모델은 최초에 장면 그래프를 검출하고 구축하기 위한 보조 모델이 필요하지 않으며, 이는 더 많은 연산 효율성을 갖는다. 그러나 전류 변환기 기반 모델은 여전히 텍스트를 위해 설계된 원래 변환기의 내부 아키텍처를 사용하는데, 각 변환계층은 단일 멀티-헤드 도트-생산 어텐션 정제 모듈을 가지고 있다. 이 구조는 이미지 영역 사이의 관계의 완전한 복잡성을 모델링할 수 없으므로, 우리는 이미지 데이터에 적응하도록 변환층 내부 아키텍처를 변경하는 것을 제안한다. 우리는 변환기 계층을 넓혀서, 각각의 변환기 계층이 인코딩 및 디코딩 단계에서 모두 상이한 영역의 양태에 대한 다수의 정제 모듈을 갖도록 한다.

## 3. Image Transformer

이 섹션에서, 우리는 먼저 원래 변환자 레이어 [11]을 검토하고, 제안된 이미지 변환기 아키텍처에 대한 인코딩 및 디코딩 부분을 정리한다.

### 3.1 Transformer Layer

트랜스퍼머는 다중 헤드 도트-Produk 어텐션 기반 트랜스퍼미니싱 레이어의 스택으로 구성된다.

Sen He 등이 연구를 수행하였다.

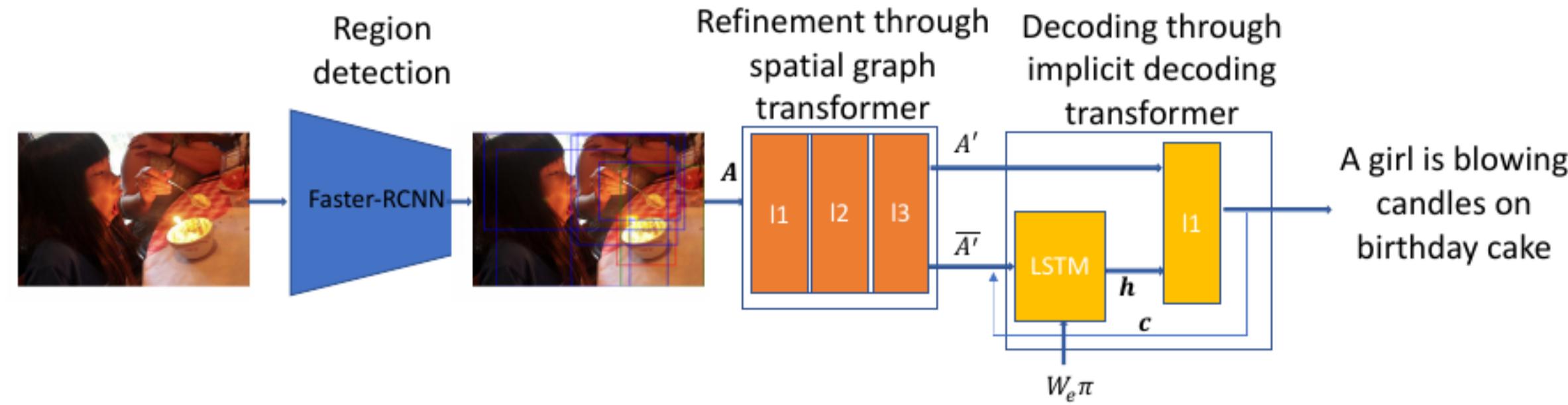


그림 1과 같이 나타났다. 2: 우리 모델의 전체 아키텍처는 정제 부분이 공간 그래프 트랜스포머 레이어를 3개 스택으로 구성하고, 디코딩 부분은 암묵 디코딩 트랜스포머 레이어를 갖는 LSTM 레이어를 갖는다.

각각의 계층에서는,  $N$  개의  $D$ -dimensional 엔트리들로 구성된, 주어진 입력  $A \in \mathbb{R}^{*}$ 에 대해, 그 계층 내에서 단계별로 처리된다. 자연어 처리에서 입력 엔트리는 문장에서 단어의 임베디드 특징(feature)이 될 수 있고, 컴퓨터 비전 또는 이미지 캡션에서 입력 엔트리는 이미지에서 영역을 설명하는 특징(feature)이 될 수 있다. 변환기의 핵심 기능은 다중-헤드 점-곱 어텐션을 통해 각 엔트리를 다른 엔트리와 함께 정제하는 것이다. 변환기의 각 레이어는 먼저 입력을 질의들( $Q = AW_g$ ,  $W_g \in \mathbb{R}^?$ )로 변환한다. \*\*), 키들( $K = AW_K$ ,  $W_K \in \mathbb{R}^{*?}$ ) \* 및 val-ues ( $V = AW_V$ ,  $W_V \in \mathbb{R}^?$ ) 선형 변환을 통해서도 비록 스케일링된 내적 주의는 다음과 같이 정의된다.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V, \quad (1)$$

여기서  $D_j$ 는 키 벡터의 차원이고  $D$ 는 값 벡터의 차원(구현 시  $D = D_j = D$ )이다. 어텐션 레이어의 성능을 향상시키기 위해 멀티 헤드 어텐션을 적용한다.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O, \\ \text{head}_i &= \text{Attention}(AW_{Q_i}, AW_{K_i}, AW_{V_i}). \end{aligned} \quad (2)$$

그런 다음, 멀티-헤드 어텐션으로부터의 출력은 입력과 함께 합산되고 정규화된다.

$$A_m = \text{Norm}(A + \text{MultiHead}(Q, K, V)), \quad (3)$$

여기서  $\text{Norm}(\cdot)$ 은 계층 정규화를 의미한다.

변환기는 각 모듈에서 잔여 연결을 구현함으로써 변환층 최종 출력은 다음과 같다.

$$A' = \text{Norm}(A_m + \phi(A_mW_f)), \quad (4)$$

비선형성을 가진 피드 포워드 네트워크(a feed-forward network)를 사용한다.

각각의 정제 계층은 이전 계층의 출력을 입력으로 취한다(제1 계층은 원래의 입력을 취한다. 디코딩 부분은 또한 변환자 정제 층들의 스택으로서, 인코딩 부분의 출력뿐만 아니라 이전 예측된 단어의 내장된 특징들을 취한다).

## tial Graph Encoding 'Transformer La



그림 3은 이미지의 영역을 표시하는 예제입니다. (a)는 이미지 내의 물체와 사람을 각각 다른 영역으로 인식하고 있는 모습입니다. (b)는 같은 이미지에서 다른 영역을 인식하는 예시로, 영역의 경계가 바뀌거나 추가되는 경우입니다.

질의와 키 쌍 사이의 공간 관계만을 간주하는 원래 변환기와 대조적으로, 우리는 인코딩 부분에서 공간 그래프 변환기를 사용할 것을 제안하고, 여기서는 그래프 구조에서 각 질의 영역에 대해 부모, 이웃, 자식(그림 1과 같은 예)을 위한 3가지 일반적인 공간 관계를 고려한다. 그리고 3). 따라서 각 층에서 3개의 서브 변환기 층을 병렬로 추가하여 각각의 변환기 층을 넓히고, 각각의 서브 변환기는 공간 관계의 카테고리를 담당하며, 모두 동일한 큐리를 공유한다. 인코딩 단계에서, 우리는 두 영역들 간의 상대적 공간 관계를 그들의 오버랩에 기초하여 정의한다. 우리는 먼저 그래프 인접 행렬  $2, \epsilon R^N * %$  (부모 노드 인접 행렬),  $2, \epsilon R \epsilon % *$  (주변 노드 인접 행렬),  $2$ 를 계산한다. 이미지의 모든 영역에 대해  $R & %$  (아동 노드 인접 매트릭스)

$$\Omega_p[l, m] = \begin{cases} 1, & \text{if } \frac{\text{Area}(l \cap m)}{\text{Area}(l)} \geq \epsilon \text{ and } \frac{\text{Area}(l \cap m)}{\text{Area}(l)} > \frac{\text{Area}(l \cap m)}{\text{Area}(m)} \\ 0, & \text{otherwise.} \end{cases}$$

$$\Omega_c[l, m] = \Omega_p[m, l]$$

$$\text{with } \sum_{i \in \{p, n, c\}} \Omega_i[l, m] = 1$$
(5)

여기서  $\epsilon = 0.9$ 인 경우는 우리 실험에서 그다지 유의미하지 않다. ‘공간 그래프 인접 행렬은 인코더에서 각 서브 변환기의 출력을 조합하기 위해 각각의 서브 변환기에 내장된 공간 하드 어텐션으로 사용된다. 보다 구체적으로는 원본의 경우에 대해 말한다.

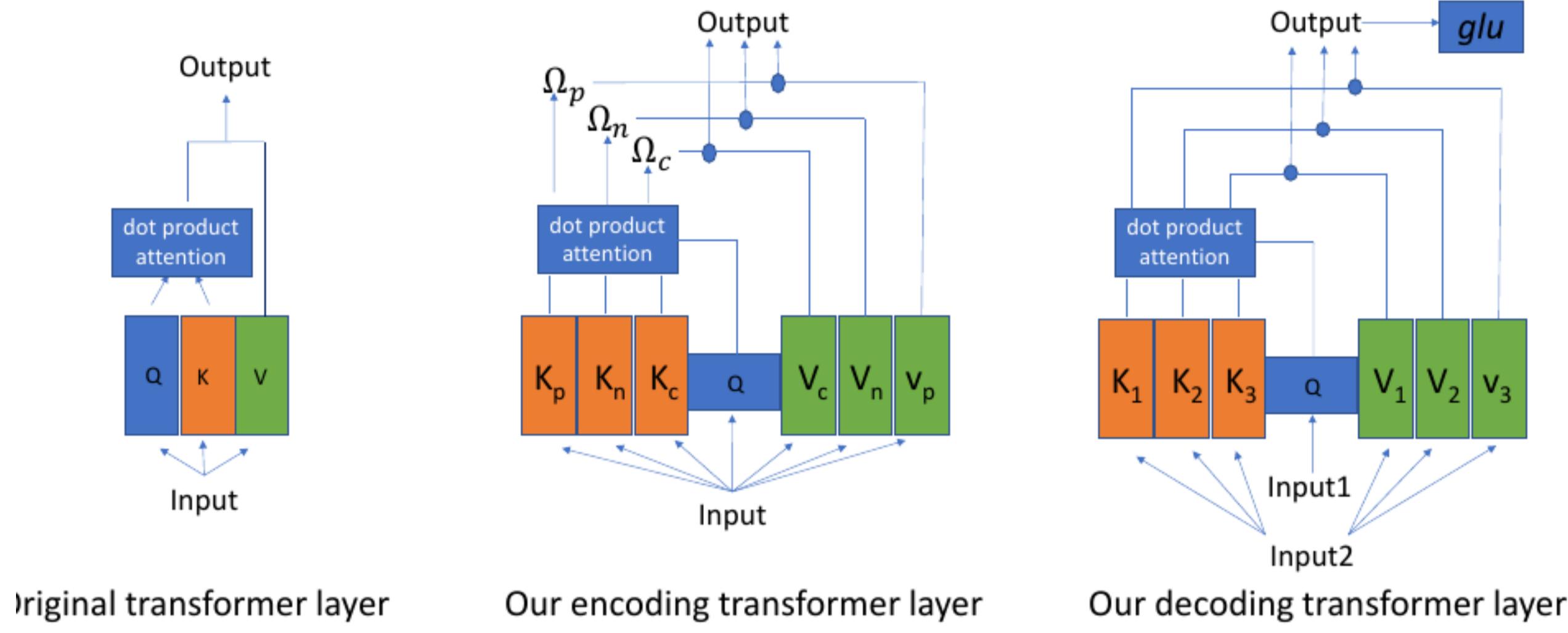


그림 1과 같은 모습을 보인다. 4: 원래의 변환기 계층과 제안된 인코딩 및 디코딩 변환기 계층들 간의 차이점이다.

이러한 인코딩 트랜스포머는 식 (2), (3), (4)에서 정의된다. (1)과 (2)는 다음과 같이 개정된다.

$$\text{Attention}(Q, K_i, V_i) = \Omega_i \circ \text{Softmax} \left( \frac{QK_i^T}{\sqrt{d}} \right) V_i, \quad (6)$$

$\circ$ 는 하다마르 곱이고,  $\circ$ 는 하다마르 곱이다.

$$A_m = \text{Norm} \left( A + \sum_{i \in \{p, n, c\}} \text{MultiHead}(Q, K_i, V_i) \right). \quad (7)$$

우리가 트랜스포머를 넓힐 때 따라, 우리는 원래의 것과 유사한 복잡성을 달성하기 위해 인코더의 스택 수를 반으로 줄인다(3개의 스택, 원래의 트랜스포머가 6개의 스택을 가진다). 우리의 공식화로, 우리는 공간 그래프와 의미 그래프를 변환기 층으로 결합했다(장면 그래프 기반 방법[7,9]은 이를 인코딩하기 위해 두 가지 가지를 필요로 한다). 상기 원래 변환기 아키텍처는 이미지 내의 어떤 영역도 다른 영역을 포함하거나 포함되지 않는다는, 제안된 아키텍처의 특별한 경우이다.

### 3.3. Implicit Decoding Transformer Layer

우리의 디코더는 LS'TM [28] 계층과 암시적 변환기 디코딩 계층으로 구성되어 있는데, 이는 이미지의 영역 내의 다양한 정보를 디코딩하기 위해 제안되었다. LSTM 레이어는 공통의 메모리 모듈이며, 트랜스포머 레이어는 도트 곱 어텐션(dot product attention)을 통해 이미지에서 가장 관련성이 높은 영역을 추론한다.

처음에 LSTM 계층은 인코딩 변환기, 마지막 시차에서의 컨텍스트 벡터( $c_{t-1}$ )와 현 문장의 현 단어의 임베디드 특징 벡터로부터 출력( $A=7$ )을 받는다.

$$\begin{aligned} x_t &= [W_e \pi_t, \bar{A} + c_{t-1}] \\ h_t, m_t &= \text{LSTM}(x_t, h_{t-1}, m_{t-1}) \end{aligned} \quad (8)$$

여기서  $W$ 는 단어 임베딩 행렬이며,  $7$ 은 실제 데이터( $t'$ )의 단어이다. 출력 상태  $h_t$ 는 이어서 선형으로 변환되고, 임플리시 디코딩 변환기 층의 입력에 대한 질의로서 처리된다. 원래 변환기 레이어와 우리의 암시적 디코딩 변환기 레이어 사이의 차이점은 각 서브 변환기가 영역의 상이한 측면을 암시적으로 디코딩할 수 있도록 하나의 레이어에서 여러 서브 변환기를 병렬로 추가하여 디코딩 변환기 레이어도 넓혀준다는 것이다. 이것은 다음과 같이 정형화되어 있다.

$$A_{t,i}^D = \text{MultiHead}(W_{DQ}h_t, W_{DKi}A', W_{DVi}A') \quad (9)$$

그런 다음 서브 변환기 출력의 평균은 채널별로 현재 단계의 새로운 컨텍스트 벡터( $c_t$ )를 추출하기 위해 게이팅된 선형 계층(GLU) [27]을 통과한다.

$$c_t = \text{GLU}\left(h_t, \frac{1}{M} \sum_{i=1}^M A_{t,i}^D\right) \quad (10)$$

그러면 맥락 벡터를 사용하여 시간 단계  $t$ 에서의 단어의 확률을 예측한다.

$$p(y_t | y_{1:t-1}) = \text{Softmax}(w_p c_t + b_p) \quad (11)$$

우리의 모델의 전체적인 구조는 그림 1에 설명되어 있다. 2, 그리고 원래 변환기 계층과 우리가 제안한 인코딩 및 디코딩 변환기 계층의 차이를 그림 2에 보여주었다. 4.

### 3.4 'Training Objectives

모델 파라미터  $\theta$ 를 학습하기 위해 목표 화가 단어  $y_j$ . $7$ 의 시퀀스로 주어지면, 우리는 이전 방법을 따르며, 우리는 먼저 크로스 엔트로피 Io로 모델을 학습한다.

$$L_{XE}(\theta) = - \sum_{t=1}^T \log(p_\theta(y_t^* | y_{1:t-1}^*)) \quad (12)$$

그 다음은 자기 비판 강화 훈련 [29] CIDEr 점수 최적화 [30]를 통해 이루어진다.

$$L_R(\theta) = -E_{(y_{1:T} \sim p_\theta)}[r(y_{1:T})] \quad (13)$$

심부 기능은 근사이며 그레디언트는 근사이다.

$$\nabla_\theta \approx -(r(y_{1:T}^s) - (\hat{y}_{1:T})) \nabla_\theta \log p_\theta(y_{1:T}^s) \quad (14)$$

## Experiment

### 4.1 Datasets and Evaluation Metrics

우리의 모델은 M5COCO 이미지 캡션 데이터 세트에 대한 학습을 수행한다. 학습 집합에 11,3287개의 이미지가 있고, 5,000개의 이미지가 있는 32]을 따르게 된다.

model	Bleu1	Bleu4	METEOR	ROUGE-L	CIDEr	SPICE
<b>single-stage model</b>						
Att2all[29]	-	34.2	26.7	55.7	114.0	-
<b>two-stages model</b>						
n-babytalk[4]	75.5	34.7	27.1	-	107.2	20.1
up-down[3]	79.8	36.3	27.7	56.9	120.1	21.4
<b>scene graph based model</b>						
GCN-LSTM*[7]	80.9	38.3	28.6	58.5	128.7	22.1
AUTO-ENC[8]	80.8	38.4	28.4	58.6	127.8	22.1
ALV*[9]	-	38.4	28.5	58.4	128.6	22.0
GCN-LSTM-HIP*†[10]	-	39.1	28.9	<b>59.2</b>	130.6	22.3
<b>transformer based model</b>						
Entangle-T*[13]	<b>81.5</b>	<b>39.9</b>	28.9	59.0	127.6	22.6
AoA[12]	80.2	38.9	<b>29.2</b>	58.8	129.8	22.4
VORN[14]	80.5	38.6	28.7	58.4	128.3	22.6
Ours	80.8	39.5	29.1	59.0	<b>130.8</b>	<b>22.8</b>

표 1: MSCOCO COCO CAPTION 오프라인 테스트 세트를 활용해 한 번에 이미지에 대한 여러 모델을 융합하는 방법을 제시합니다.

검증 세트와 테스트 세트의 5,000개의 이미지이다. 각각의 이미지는 5개의 캡션을 그라운드 트루스로서 갖는다. 4번 미만으로 나타나는 단어들을 버리고, 최종 어휘 크기는 10,369개이다. 우리는 카르파티의 오프라인 테스트 세트(5,000개의 이미지)와 MSCOCO 온라인 테스트 데이터세트(40,775개의 이미지)에서 모델을 테스트한다. 평가 메트릭은 Bleu [33], METEOR [34], ROUGE-L [35], CIDEr [30], SPICE [23]를 사용한다.

## 4.2 Implementation Details

이전 작업에 따라, 우리는 먼저 Faster R-CNN를 Visual Genome [21]에 대해 트레이닝하고, ImageNet [16]에 대해 미리 트레이닝한 resnet-101 [36]을 백Bone으로 사용한다. 각각의 이미지에서 우리는 10~100개의 정보적인 영역을 검출할 수 있는데, 각 영역의 경계는 먼저 정규화되고 그 후 공간 그래프 행렬을 계산하는데 사용된다. 그런 다음, 우리는 각 이미지 영역에 대한 추출된 특징과 계산된 공간 그래프 매트릭스를 이용하여 이미지 캡션화를 위한 제안 모델을 훈련시킨다. 우리는 먼저 25개의 에폭을 통해 교차 엔트로피 손실로 모델을 훈련하며, 초기 학습률은  $2 \times 10^{-3}$ 으로 설정하고, 3개의 에폭마다 학습률을 0.8씩 감소시킨다. 우리의 모델은 10개의 배치 크기를 갖는 Adam[37]을 통해 최적화된다. 그런 다음, 우리는 또 다른 35개의 에포크 동안 강화 학습을 통해 우리의 모델을 더욱 최적화한다. 디코더의 LSTM 층의 크기는 1024로 설정되었으며, 추론 단계에서 크기 3의 빔 탐색이 사용된다.

## 4.3. Experiment Results

우리의 모델의 성능을 발표된 이미지 캡션 모델과 비교한다. 비교된 모델에는 최상위 성능의 단일 스테이지 어텐션 모델, Att2all[29], 두 단계 어텐션 기반 모델인 n-아비토크[4]과 up-down[3], 비주얼 장면 그래프 기반 모델인 GCN-LSTM[7], AUTO-ENC[8], ALV[9], GCN-LSTM-HIP[10], 트랜스포머 기반 모델 Entangle-T[13], AoA가 포함된다.

model	B1		B4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
<b>scene graph based model</b>										
GCN-LSTM*[7]	80.8	95.9	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
AUTO-ENC*[8]	-	-	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
ALV*[9]	79.9	94.7	37.4	68.3	28.2	37.1	57.9	72.8	123.1	125.5
GCN-LSTM-HIP*†[10]	81.6	95.9	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
<b>transformer based model</b>										
Entangle-T*[13]	81.2	95.0	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
AoA[12]	81.0	95.0	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
Ours	81.2	95.4	39.6	71.5	29.1	38.4	59.2	74.5	127.4	129.6

표 2: MSCOCO 온라인 테스트 서버에서 최근 발표된 모델들의 리더보드이다. \*은 두 개의 모델의 융합을 의미한다. \*은 특징 추출 골격으로 SENet [31]을 의미한다.

이것은 {12}, VORN [14]의 결론으로, 그들이 하였다. M5COCO Karpathy 오프라인 테스트 세트에 대한 비교는 표 1에 설명되어 있다. 우리의 모델은 CIDEr 및 SPICE 점수에서 새로운 최고의 상태를 달성하는 반면, 다른 평가 점수는 이전의 최상위 성능 모델과 비교할 수 있다. 주목할 것은 대부분의 시각적 장면 그래프 기반 모델이 의미론적 및 공간적 장면 그래프를 융합하였고, 보조 모델이 초기에 장면 그래프를 구축하도록 요구했기 때문에, 우리의 모델이 보다 연산적으로 효율적이라는 점이다. VORN[14]도 공간 주의를 자신의 모델에 통합했으며, 모든 종류의 평가 메트릭 중 우리의 모델이 더 잘 수행하고, 이것은 우리의 공간 그래프 변환기층이 우수함을 보여준다. MSCOCO 온라인 테스트 결과는 표에 나와 있다. 2, 우리의 모델은 여러 평가 메트릭에서 이전 트랜스포머 기반 모델보다 뛰어나다.

#### 4.4 Ablation Study and Analysis

절제 연구에서, 우리는 AoA[12]를 강한 기준선 °(레이어당 단일 멀티-헤드 도트-프로덕션 어텐션 모듈이 있는)로서 사용하며, 이는 멀티-헤드 어텐션 위에 게이트된 선형 레이어[27]를 추가한다. 인코더 부분에서는 Eqs를 개정하여 각 계층의 세 개의 하위 변환기의 평균 출력을 취하여 공간 관계를 제거하는 인코더에서 공간 관계의 영향을 연구한다. 6과 7은: Attention(Q, K; V) = Softmax (4S) V; Am = Vd Norm(4+3 Vietp,ne) MultiHead(Q, Kj, Vi): 우리는 또한 제안된 공간 그래프 인코딩 변환층을 인코딩 부분에서 어디에 사용할 것인가? 1층, 2층, 3층 또는 이들 중 세 개를 사용하는지 연구한다. 복호화 부분에서는 서브 변환기의 개수(M 등 식 2)가 미치는 영향을 연구한다. 암시적 디코딩 변환층에서 10)과 같이 할 수 있다.

그림에서 알 수 있듯이, 우리는 이와 같은 내용을 확인할 수 있다. 3, 부호화 변환 계층을 넓혀 모델의 성능을 크게 향상시켰다. 인코딩 트랜스포머의 모든 계층이 동일하지 않지만, 인코딩 부분의 최상위 계층에 제안한 트랜스포머 계층을 사용할 때, 개선점이 감소되었다. 이것은요.

~ Our experiments are based on the code released at: <https://github.com/husthuaan/AoANet>

model	Bleu1	Bleu4	METEOR	ROUGE-L	CIDEr	SPICE
baseline(AoA)	77.0	36.5	28.1	57.1	116.6	21.3
<b>positions to embed our spatial graph encoding transformer layer</b>						
baseline+layer1	77.8	36.8	28.3	57.3	118.1	21.3
baseline+layer2	77.2	36.8	28.3	57.3	118.2	21.3
baseline+layer3	77.0	37.0	28.2	57.1	117.3	21.2
baseline+layer1,2,3	77.5	37.0	28.3	57.2	118.2	21.4
<b>effect of spatial relationships in the encoder</b>						
baseline+layer1,2,3 w/o spatial rela	77.5	36.8	28.2	57.1	117.8	21.4
<b>number of sub-transformers in the implicit decoding transformer layer</b>						
baseline+layer1,2,3 (M=2)	77.5	37.6	28.4	57.4	118.8	21.3
baseline+layer1,2,3 (M=3)	78.0	37.4	28.4	57.6	119.1	21.6
baseline+layer1,2,3 (M=4)	77.5	37.8	28.4	57.5	118.6	21.4

Table 첫째줄은 규칙에 따른 단위별로 정리된 결과이다. 세 번째 줄은 각 단위별로 연산망과 학습 번째  
계층 단위별로 각 단위별로 연산망을 예상해놓은 원래의 것을 원형함수를 활용한 링 범용자체망의 계층  
변환 가수와 비교하여 대개수이다.

변환기의 최상위 계층에서의 공간 관계가 그만큼 정보가 없기 때문일 수 있지만, 우리는 인코딩 파트에서 모든 계층에서 우리의 공간 변환 계층을 사용한다. 제안한 보다 넓은 변환기 층에서 공간 관계를 줄일 때, 그에 따른 성능 감소도 일부 존재하는데, 이는 우리의 설계에서 공간 관계의 중요성을 보여준다. 복호화 변환기를 확대한 후에 향상도를 더욱 증가시켰고 (복호화 변환기 레이어는 3개의 서브 변환기로 확대한 후에 CIDEr 점수는 118.2에서 119.1로 증가) 더 넓혀서는 더 좋은 결과를 얻을 수 있지만 복호화 변환기 레이어에 4개의 서브 변환기가 있는 경우에는 성능이 약간 떨어지므로 최종적으로 우리의 복호화 변환기 레이어는 3개의 서브 변환기가 병렬로 우리의 모델 결과의 정성적인 예는 그림에 예시되어 있다. 다섯 번째는 5번이죠. 우리가 볼 수 있듯이, 공간 관계를 포함하지 않은 기준 모델은 붉은 버스 위의 경찰관(오른쪽 위), 그리고 기차 위의 사람(왼쪽 아래)을 잘못 묘사했다.

암묵적 그래프 시각화 인코딩: 변환기층은 암묵적 그래프로 간주될 수 있으며, 이는 도트-Produk 어텐션을 통해 정보 영역들을 연관시킨다. 여기서 우리는 도구를 통해 우리가 제안한 공간 그래프 변환층이 주의를 통해 정보 영역을 연결하는 방법을 시각화한다. 6. 첨단 예에서는 원래 변환계층이 기차와 산에 있는 사람들을 강하게 연관시키고 잘못된 설명을 내놓는 반면, 우리가 제안한 변환계층은 기차와 선로, 산과 연관시키고, 원래 변환계층은 곰과 물에 비친 모습을 ‘두 곰’으로 다루는 반면, 우리 변환계층은 곰과 물에 비친 모습을 구별하고 눈에 달린 곳으로 연관시킨다.

디코딩 특징 공간 시각화: 우리는 또한 디코딩 변환 계층 출력을 시각화했다. 7)은 그만큼 더 많은 수의 사례를 보여준다. 내부에 하나의 서브 변환기만을 갖는 원래의 디코딩 변환기 층과 비교된다. 우리가 제안한 암시적 디코딩 변환기 층의 출력은 축소된 특징에서 더 큰 영역을 커버한다.

## Image Captioning through Image Transformer 13



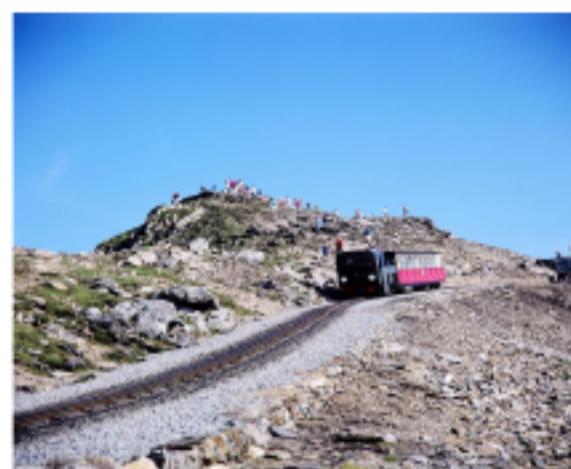
높은 건물들로 둘러싸인 거리의 신호등.  
높은 초고층 건물이 있는 도시를 흑백으로  
찍은 사진.  
어떤 건물들은 신호등과 흐린 하늘을  
보이고 있다.  
\* 거리에서 찍은 신호등 흑백  
사진  
\* 건물로 둘러싸인 신호등과  
가로표지판.

기준선: 도시의 한 거리에 있는 교통등 두 개다.  
우리들은: 건물이 있는 거리의 신호등이다.



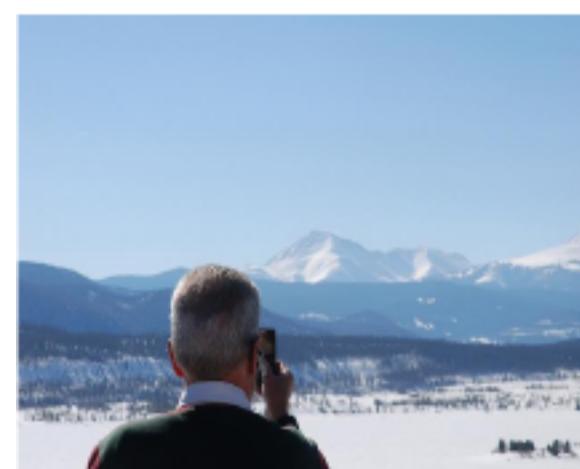
GT  
우 빨간 버스 앞에 서 있는 경찰관들.  
세 명의 자전거 운전자가 거리에서  
빨간 버스와 마주쳤다.  
« 어떤 사람들은 오토바이에서 빨간  
버스를 달고 있다.  
바이크를 타고 있는 남성 몇 명이  
빨간색 버스를 지나고 있다.  
그 옆으로 주차관리원들이 버스와  
나란히 타고 있다.

기준선: 적색 버스에 타고 있는 경찰관 그룹. 우리들은: 빨간 버스 앞에  
모터사이클을 타고 다니는 경찰관들 그룹.



선로를 따라 내려가는 기차의 모습을  
담아냅니다.  
\* 철로가 있는 바위에 몇몇 사람들이 서  
있어 있다.  
아트레인은 낮에 언덕 위에서 선로를  
따라 이동한다.  
† 열차 차량이 한 언덕 위에서 선로를  
과한다.

기준선: 선로 위의 기차를 타고 있는 사람들의 그룹이다.  
우리들은: 기차가 산 위의 선로를 따라 내려가고 있다.



GT  
아만은 산 앞에서 휴대전화를 들고  
있다.  
눈이 깔린 비탈길 위에 서 있는 나이 듦  
남자.  
아만이 넓은 산세를 바라보는 모습.

기준선: 한 남성이 산 위에서 휴대폰을 쥐고 있다. 우리들은: 한 남성이 휴대폰으로 산을  
찍고 있다.

그림 1은 공기 중의 오염 물질과 수소 분자 간의 물리적 상호 작용을 나타내고 있다. 5: MSCOCO 이미지 캡션 데이터셋트 [17]에서 우리의 방법에 대한 질적 예시를 사용하여 지상 진단 및 강력한 베이스 라인 방법(AoA [12])과 비교하였다.

원래보다 많은 공간을 의미하며, 이는 이미지 영역에서 더 많은 정보를 디코딩하는 것을 의미한다. 복호화 변환층의 출력으로부터의 원래 특징 공간(1,024차원)에서, 우리는 1,000개의 예로부터의 특징 맵의 공분산 행렬의 트레이스를 계산하는데, 원래 변환층의 트레이스는 30.40이고, 우리의 더 넓은 복호화 변환층의 트레이스는 454.57로, 이는 우리의 설계가 특징 공간에서 더 큰 영역을 커버하는 디코더의 출력을 가능케 한다는 것을 나타낸다. 그러나 복호화 변환기 층에서 개별 서브 변환기의 경우 여전히 특징 공간에서 다른 요인들을 분리하는 것을 학습하지 못하는 것처럼 보이며 (각 서브 변환기의 출력으로부터 뚜렷한 클러스터가 없기 때문에), 우리는 이들이 출력에 대한 직접적인 감독이 없기 때문에 분리된 특징을 자동으로 학습할 수 없을 수 있다.

## 5 Discussion and Conclusion

본 작업에서는 timage 변환 아키텍처를 소개하였다. 제안된 아키텍처의 핵심 아이디어는 기계 번역을 위해 설계된 원래 변환기 레이어를 이미지의 구조에 적응하도록 넓히는 것이다. 인코더에서는 이미지 영역들 사이의 공간 관계를 이용하여 변환기 레이어를 넓히고, 디코더에서는 더 넓은 변환기 레이어가 이미지 영역들에서 더 많은 정보를 디코딩할 수 있다. 제안된 모델의 우수성을 보여주기 위해 광범위한 실험이 수행되었고, 제안된 인코딩 및 디코딩 변환 계층의 유효성을 검증하기 위해 실험에서 정성적 및 정량적 분석이 설명된다. 이전 이미지 캡션에서 최고 모델에 비해, 우리 모델은 새로운 최첨단 SPICE 스코어를 달성하는 반면, 이 중에서 가장 높은 스코어를 달성한다.



그럼 1운자기형영광주위부와 학생들은 물론 평가형겠지. 또 6의 차라 형역에 꼬주의지를 활용해 세 가지의 차기형 청 생역을 차기형성기세포와 판형성사체포와 함께 세포와 원당격형성 세포와 역할을 가짐. 차기형역이 전기형 청 생역을 차기형성세포와 자기형성세포의 영역인 형성포와 자카 형성세포와 차기형성세포와 자카 형성세포를 차기형역이나 주위를 통해 서로 다른 키 영역들고 여전히 향수의 차를 차고 있고 바운딩 밤수나의 영역은 차기형역이나 주위를 통해 서로 다른 키 영역들고 여전히 향수의 차를 차고 있고 바운딩 밤수나의 영역은 차리 영역이고 다른 영역들은 키 영역이다. 각각의 키 영역의 투명성은 질의 영역과의 도트-PRODUCT 주의 가중치를 보여준다. 투명성이 높다는 것은 더 큰 도트-공간 주의 가중치를 의미하며, 그 반대이다.

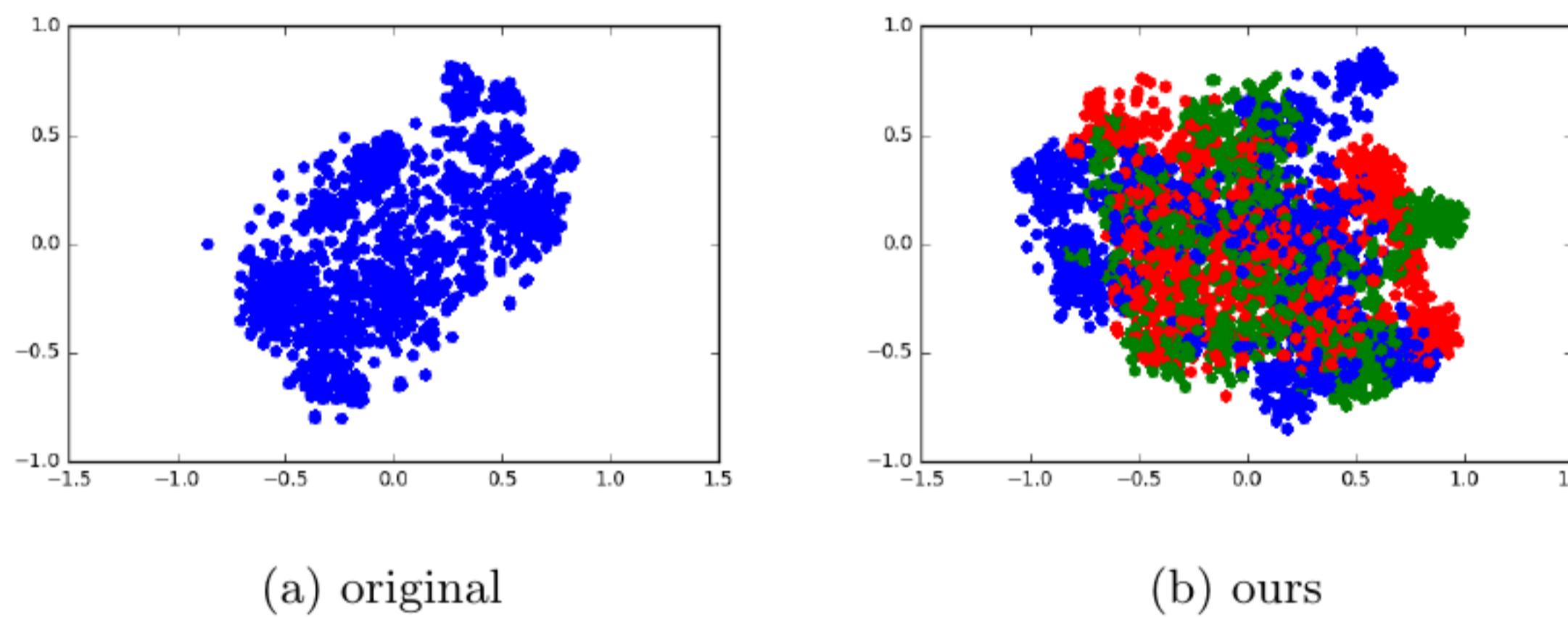


그림 1은 고형물질의 형태를 나타내는 것이다. 7: t-SNE [39] 디코딩 트랜스포머 계층에서 출력(1,000 예) 시각화, 상이한 컬러는 우리 모델의 디코더에서 상이한 서브-트랜스포머로부터 출력을 나타낸다.

다른 평가 메트릭, 우리의 모델은 비교 가능하거나 이전의 최고 모델을 능가하며, 더 나은 계산 효율성을 갖는다.

우리의 작업은 커뮤니티가 이미지 캡션뿐만 아니라 그 내부의 관계적 주의가 필요한 다른 컴퓨터 비전 작업에 도움이 될 수 있는 더 발전된 변환기 기반 아키텍처를 개발하도록 영감을 불러일으키기를 바란다. 우리의 코드는 향후 연구를 지원하기 위해 커뮤니티에 공유될 것이다.

# Image Captioning through Image Transformer 1

## References

17번. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L. MS 코코 캡션: 데이터 수집 및 평가 서버. arXiv 사전 프린트 arXiv:1504.00325 (2015)

18번이 되면 18번이 된다. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. IEEE 컴퓨터 비전 및 패턴 인식 학회 프로시던스 (2015) 3156-3164)에서 정의한 것과 동일한 개념을 사용하였다.

19. Gers, F.A., Schmidhuber, J., Cummins, F.: 잊기 학습: lstm을 이용한 지속적 예측. 신경 계산 12 (2000) 2451-2471

20번이요. Lu, J., Xiong, C., Parikh, D., Socher, R.: 언제 볼 것인지 아는 것: 이미지 캡션을 위한 시각적 감시원을 통한 적응적 주의를 통해. : IEEE의 컴퓨터 비전과 패턴 인식에 관한 회의의 공장물이다. 이것은 (2017) 375-383에서 정의한 바와 같다.

21번이요. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A. 등. : 시각적 개념: 크라우드 소싱된 조밀한 이미지 주석을 사용하여 언어와 시각을 연결한다. 국제 컴퓨터 비전 저널 International Journal of Computer Vision 123 (2017) 32-73

22번입니다. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L. MS 코코 : 맥락에 있는 공통의 객체이다. European conference on computer vision, Springer (2014) 740-755.

23번이요. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: 시맨틱 명제 이미지 캡션 평가. 유럽 컴퓨터 비전 학회, 스프링어(2016) 382-398.

24번입니다. Hao, J., Wang, X., Shi, S., Zhang, J., Tu, Z.: 뉴럴 기계 번역을 위한 다중 그랜불러리 셀프-어텐션. 아르씽비프리프린트 arXiv:1909.02222 (2019)

25개다. 王, X., Tu, Z., Wang, L., Shi, S.: 구조적 포지션 표현을 갖는 자기 주의이다. 아르빅 사전 프린트 arXiv:1909.00383 (2019)

26번입니다. Wang, Y.S., 이, H.Y., Chen, Y.N. Tree 변환기 : 나무 구조를 셀프-어텐션에 통합한다. 아르시브 사전 프린트 arXiv:1909.06639(2019)에 실렸다.

27. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional network. In: 34차 국제 기계학습 콘퍼런스-70권- JMLR (2017) 933-941

28개이다. Hochreiter, S., Schmidhuber, J.: Long short-term memory. 신경 계산 9(1997) 1735-1780

29. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: 이미지 캡션을 위한 자기 비판적 시퀀스 훈련. In: IEEE 컴퓨터 비전 및 패턴 인식 회의의 프로시저. (2017) 7008-7024번의 경우는 위와 같은 결과를 얻을 수 있다.

30번. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-Based Image Description 평가. In: IEEE 컴퓨터 비전 및 패턴 인식에 관한 회의의 프로시저스. (2015) 4566-4575로 하여금 이에 대한 논의를 하게 하였다.

31번이요. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation 네트워크(squeeze-and-excitation network). IEEE 컴퓨터 비전 및 패턴 인식 학회 회의의 발표 결과에서. (2018) 7132-7141호로 하여금, 본 고의 수용권을 수용권자에게 통제할 수 있다.

32번입니다. 카르파시, A., 피-페이, L.: 이미지 설명을 생성하기 위한 심층 시각-시맨틱 정렬이다. Computer Vision 및 패턴 인식에 관한 IEEE 회의의 논문입니다. (2015) 3128-3137번으로 제안된 바와 같이, 그에 대한 논의가 진행되었다.

33번입니다. Papineni, K., Roukos, S., Ward, T., Zhu, W.J. Bleu : 기계 번역의 자동 평가를 위한 방법이다. In: 연산 언어학 협회 연례 회의 40회(2002) 연산 언어학 협회 311-318

34번이 있습니다. Banerjee, S., Lavie, A.: Meteor: 인간 판단과 향상된 상관관계를 가진 MT 평가를 위한 자동 메트릭이다. In: 기계 번역 및/또는 요약을 위한 내재적 및 외재적 평가 측정치에 대한 acl 워크숍의 프로시더스. (2005) 65-72로 정리한 바 있다.

## Image Captioning through Image Transformer 17

35번이죠. Lin, C.Y. : Rouge : 요약 자동 평가를 위한 패키지이다. In: Proc. ACL 워크숍이 텍스트 요약에 관한 분과로 발전한다. (2004) 10은 한편으로는 이와 같은 연구 결과를 통제하고자 한 것이다.

36번입니다. 그는 이미지 인식을 위한 깊은 잔류 학습을 수행했다. In: IEEE 컴퓨터 비전 및 패턴 인식 콘퍼런스의 프로시드. (2016) 770-778)에서 이를 확인할 수 있다.

37번입니다. Kingma, D.P., Ba, J.: Adam: 확률 최적화 방법. arXiv 프리프린트 arXiv:1412.6980 (2014)

38번입니다. Locatello, F., Bauer, S., Lucic, M., Ratsch, G., Gelly, S., Schdlkopf, B., Bachem, O.: 뭉쳐지지 않은 표현의 비지도 학습에서 공통 가정에 도전한다. In: 제36차 국제 기계 학습 컨퍼런스-Proceedings of the 36th International Conference on Machine Learning-Volume 97, JMLR(2019) 4114—4124

그리고 39명이 있습니다. Maaten, L.v.d., Hinton, G.: t-sne을 이용한 데이터 시각화. 머신 러닝 연구 저널 9 (2008) 2579-2605