ETRI

이미지 캡션 기술 동향

Trends On Image Captioning

하유진 20110629 |Y.J. Ha, dbwls9649@gmail.com| 세종대학교 컴퓨터 공학과

Abstract

시각(Vision)과 언어(Language)를 연결하는 것은 Generative Intelligence 에서 필수적인 역할을 하고 있고 최근 많은 사람들이 관심을 가지고 있는 연구이다. 이러한 이유로 구문론적 그리고 의미론적으로 이미지의 내용을 설명하기 위해 올바른 문장을 생성하는 이미지 캡션에 관한 여러 연구가 진행되었다. 2015 년부터 이미지 캡션 작업은 일반적으로 텍스트 생성을 위한 시각적 인코더와 문장 생성을 위한 언어 모델로 구성된 디코더 구조로 해결되었다. 이 기간 동안 두 구성 요소는 개체 영역, 속성의 활용, multi-modal 연결의 도입, fully-attentive 접근 방식, BERT 와 혯와 같은 사전 학습 언어 모델과의 융합 전략을 통해 상당히 발전했다. 그러나 이러한 결과에도 불구하고 이미지 캡션에 대한 연구는 아직 결정적인 해결책에 도달하지 못했다. 본 분석에서는 시각적 인코딩 및 텍스트 생성 방식에 관한 여러 이미지 캡션 접근 방식의 포괄적인 동향을 제공하는 것을 목표로 한다. 이와 관련하여 동향 분석에서는 아키텍처 및 훈련 전략에서 가장 영향력 있는 모델을 식별하기 위해 많은 최신 접근 방식을 정량적으로 비교한다. 이 작업의 최종 목표는 이미지 캡션 연구에 관한 기존 연구를 이해하고 앞으로의 방향을 찾기 위한 역할을 한다.

I. 서론

이미지 캡션은 이미지의 시각적 내용을 자연어로 설명하는 Vision-Language 작업 중 하나이며 기계 번역 문제에서 영감을 얻었다. 이미지 캡션 기술은 다양한 시나리오에서 활용될 수 있다. 드라마와 예능과 같은 영상에서 자동으로 자막을 생성할 수 있고 데이터 라벨링 과정에서도 자주 쓰인다. 또한, 이미지를 텍스트로 바꾸는 AI 서비스에서 활용되고 시력이 좋지 않은 사람들에게 주변 환경, 인물, 그리고 사물 등을 음성으로 묘사해 더 정확하고 풍부한 정보를 전달하여 시력 약자들의 주변 상황 이해를 도와준다.

일반적으로, 이미지 캡션 모델은 인코더-디코더 구조를 가진다. 먼저, 입력 이미지를 feature vector로 변환하는 인코딩 과정을 거친 후 이 feature vector는 문장 생성을 위한 디코더에 공급된다. 초기의 연구는 GoogleNet[1], AlexNet[2], VGG16[3]과 같은 사전 학습한 분류 네트워크를 활용하여 이미지에서 global feature를 추출했다. 그러나 이러한 feature는 정보의 과도한 압축으로 이어지는 단점이 존재한다. 이후에는 visual cues를 활용하기 위해 attention 메커니즘이 사용되어 특정 visual features에 초점을 맞춘다. 후기의 연구는 object detection network를 사용해 이미지에서 더 많은 표현을 가진 feature를 추출했다. 특히 ImageNet[4]과 Visual Genome[5]에 대해 사전 학습된 Faster-RCNN이 object detector[6]가 주로 사용되었다. 사전 학습된 object detection network는 유명한 COCO benchmark[7]에는 사용 할 수 있지만 다른 새로운 데이터에서 사용하려면 추가적인 object detection annotations가 요구되는 단점이 존재한다. 최근 모델들은 self-attention을 적용하거나 표현력이 뛰어난 visual Transformer를 인코더로 사용한다. OpenAI에서 공개한 CLIP[8]의 image encoder를 사용해 이미지에서 표현력이 풍부한 임베딩을 추출할 수 있다.

모델은 캡션을 생성하기 위해 textual decoder가 사용된다. 초기의 연구는 LSTM 변형 네트워크를 사용했다. 이러한 LSTM 종류의 네트워크는 순차적은 특성을 가지고 있어서 계산이 느리며 병렬 처리가 불가능한 단점이 존재한다. 반면 후기의 연구는 개선된 Transformer[10] 기반의 모델을 사용한다. 최근에는 Transformer를 기반으로 구축된 BERT[11], GPT[12]와 같은 사전 학습된 언어 모델이 디코더로 주로 사용된다.

다른 대규모 작업에 대해 사전 학습된 개별의 vision 모델과 언어 모델을 사용하며 훈련의 일부로 grounding을 학습한다. 따라서 쌍으로 구성된 visiolinguistic 데이터가 제한되거나 평향될 때 종종 제대로 일반화되지 않

는 myopic groundings이 생긴다. vision-and-language 작업에 대한 pretrain-then-transfer 학습 접근법은 컴퓨터 비전과 자연어 처리에서 널리 사용되기 때문에 자연스럽게 따르게 된다.

본고에서는 앞서 언급했듯이, 이미지 캡셔닝 모델의 발전 동향을 다룰 것이다. Visual encoding 부분과 textual decoding 부분을 따로 살펴 볼 것이다.

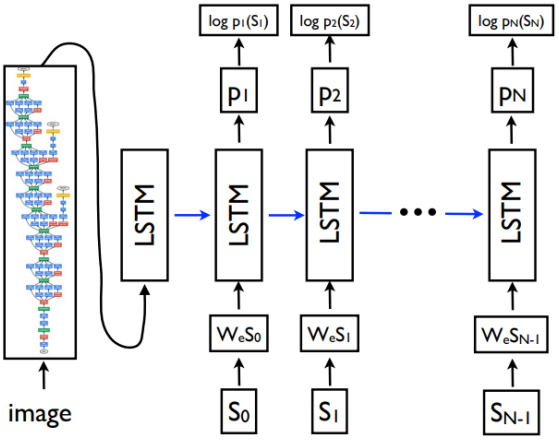
II. 이미지 캡셔닝 기술 동향

1. visual encoder

1절에서는 이미지 캡션 구조의 visual encoder의 발전 동향에 대해 다루고자 한다. 이미지에서 feature를 추출하기 위해 사전 학습된 분류 모델을 적용한 것부터 더 많은 표현을 가지는 feature를 추출하기 위해 사용한 object detection network 그리고 대규모 데이터에 대해 사전 학습된 CLIP의 이미지 인코더를 사용하기 까지의 발전 과정과 각 방법의 장단점을 소개한다.

가. 사전 학습된 분류 네트워크

CNN의 등장으로 시각적 입력을 사용하는 모든 모델의 성능이 향상되었다. 이미지 캡션의 시각적 인코딩 단계도 예외가 아니다. 가장 간단한 방법으로 CNN의 마지막 레이어 중



(그림 1) Show and Tell[13]의 아키텍처

하나를 활성화하여 고수준 및 고정 크기의 표현을 추출하여 언어 모델의 조건부 요소로 사용하는 것이다. 이는 "Show and Tell"[13]에서 사용된 방식이며 ImageNet 데이터에 대해 사전 학습된 GoogleNet을 사용하여 이미지에서 global feature를 추출한다. 추출된 feature는 언어 모델의 초기 은닉 상태에 공급된다. 같은 해 Karpathy et al.[14]는 AlexNet에서 추출한 global feature를 언어 모델의 입력으로 사용했다. 또한 Mao et al.[15] 및 Donahue et al.[16]은 VGG 네트워크에서 추출된 global features를 언어 모델의 각 time-step에서 공급했다.

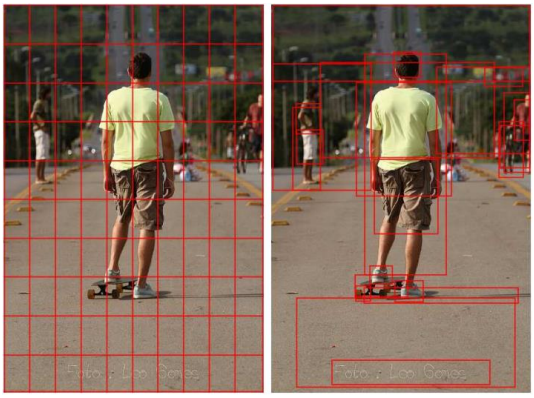
나. Attention 메커니즘

Global CNN features 사용의 주요 장점은 representation의 단순성과 간결함에 있으며 이는 전체 입력으로부터 정보를 추출하고

압축하는 능력을 가지며 이미지의 전반적인 context를 고려한다. 그러나 이러한 패러다임은 정보의 과도한 압축으로 이어지고 세분성이 부족하다. 모든 두드러진 개체와 영역이 단일 벡터에 융합되어 캡션 모델이 구체적이고 세밀한 설명을 생성하기가 어렵다. 이러한 global 표현의 단점에 자극을 받아, 이후의 접근 방식은 시각적 인코딩의 세분화 수준을 높였다 [23],[24]. 기계 번역에서 파생된 추가적인 attention 메커니즘은 광범위한 작업에서 놀라운 성능을 입증했으며 시간에 따라 변하는 시각적 feature 인코딩을 이미지 캡션 아키텍처에 부여해 더 큰 유연성과 세분성을 가능하게 한다.

다. Object Detection Network

지금까지 언급된 캡션 모델은 모두 top-down 방식으로 이미지를 이해한다. Top-down 방식은 우리의 지식과 귀납적 편견을 활용해 다가오는 감각 입력을 예측하는 것으로 구성된다. 반면 bottom-up 방식은 입력 신호에서 해석으로 전달하여 이전의 예측을 조정하는 시각적 자극을 지속적으로 제공한다. Anderson et al.[6]은 top-down 방식에 이미지 영역 제안을 담당하는 object detector로 정의되는 추가적인 bottom-up 방식을 결합하는 해결책을 제시한다. 이러한 접근법에서 Region Proposal Network가 CNN의 중간



(그림 2) top-down(left) 및 bottom-up(right)

[출처] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

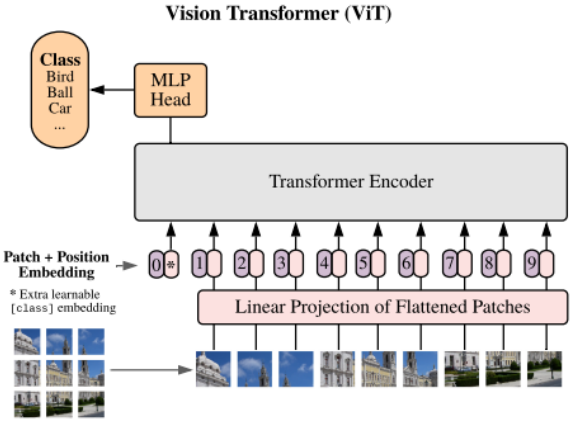
feature에 걸쳐 객체 proposal을 생성한다. 두 번째는 각 proposal에 대한 feature vector를 추출하기 위해 관심 영역의 출링을 작동한다. 이 접근법의 핵심 요소는 사전 학습 전략에 있다. 여기서 Visual Genome 데이터에서 객체 클래스와 함께 속성 클래스를 예측하는 방법을 학습하기 위한 보조 훈련 손실이 추가된다. 이를 통해 모델은 두드러진 개체와 contextual 영역 모두를 포함하고 더 나은 feature 표현을 학습하는 것을 선호하며 밀도가 높고 풍부한 예측을 할 수 있다. 제안된 모델을 통해 COCO benchmark에서 추출된 bottom-up feature vecotrs는 최근 많은 연구에서 기반으로 한다[19]-[22].

라. Slef-Attention Encoding

Self-attention은 집합의 각 요소가 다른 모든 요소와 연결되는 attentive 메커니즘이며 이는 residual connection을 통해 동일한 요소 집합의 정제된 표현을 계산하기 위해 채택된다. Yang et al. [25]은 object detector에서 비롯된 features 간의 관계를 인코딩하기 위해 self-attentive module을 사용했다. Herdade et al. [19]은 영역 간의 공간적 관계를 고려한 수정된 버전의 self-attention을 도입했다. 특히, 객체 쌍 사이의 추가적인 geometric 가중치가 계산되고 attention 가중치를 스케일하는 데 사용된다. Huang et al. [20]은 "Attention on Attention"이라는 attention 연산자의 확장을 제안했는데, 여기서 최종 attended 정보는 최종으로 함께 곱해지는 정보 벡터와 게이트 벡터를 계산함으로써 가중치가 부여된다.

마. Vision Transformer

Transformer와 같은 아키텍처는 이미지 패치에 직접 적용할 수도 있으므로 컨볼루션 연산자의 사용을 제외하거나 제한할 수 있다 [26]. Liu et al.[27]은 이미지 캡션에 대해 최초로 convolution을 사용하지 않는 아키텍처를 고안했다. 구체적으로 사전 훈련된 Transformer 네트워크(i.e. ViT[26])가 인코더로 채택되고 캡션을 생성하기 위해 표준

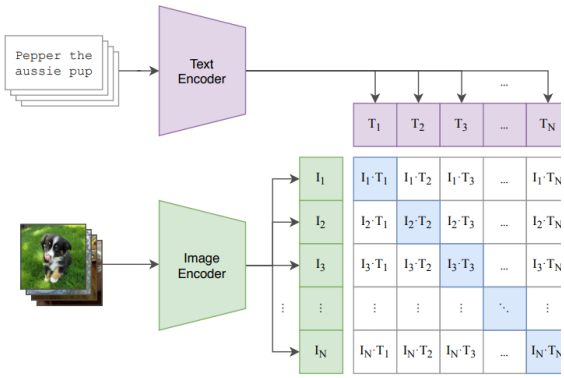


(그림 3) ViT [26]

Transformer 디코더가 사용된다.

바. CLIP

OpenAI에서 제안한 CLIP은 기존 ImageNet 보다 방대한 4억 개의 이미지 데이터를 사용해 representational learning을 수행한다. CLIP은 (이미지, 물체 라벨) 데이터 대신 (이미지, 텍스트) 데이터를 사용한다. 그러므로 분류 문제로 학습하는 것이 아닌 N개의 이미지와 N개의 텍스트 간의 올바른 연결 관계를 찾는 문제로 네트워크를 학습한다. Transformer 기반의 이미지 인코더와 텍스트 인코더가 존재하며 각 인코더를 통과해서 나온 N개의 이미지와 텍스트 feature vector들 간의 올바른 연결 관계를 학습한다. Mokady et al. [28]은 vision-language 사전 학습 모델인 CLIP 모델의 시각적 인코더를 사용하여 이미지에서 feature를 추출하고 이미지에 대한 캡션을 생성하기 위해 GPT2를 사용한다.



(그림 4) CLIP 구조 [8]

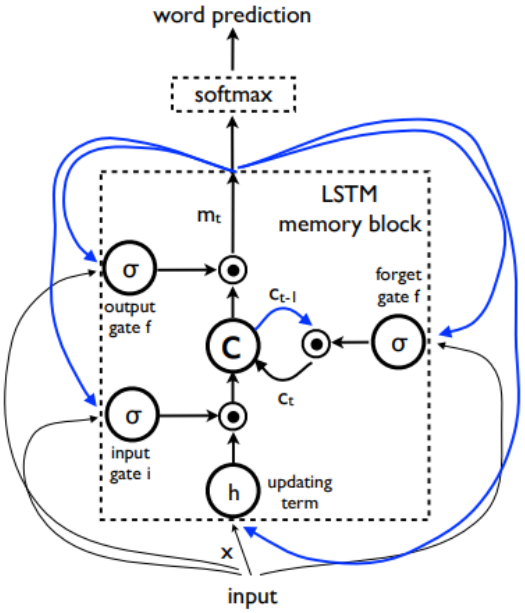
CLIP 임베딩을 GPT2 공간으로 변환하는 매핑 네트워크인 MLP를 훈련한다.

2. Textual decoder

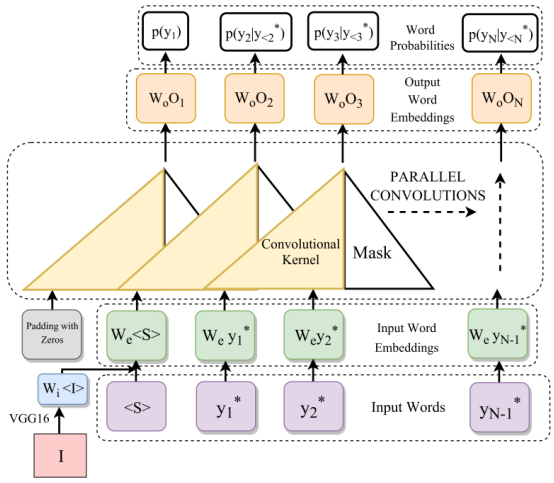
2절에서는 이미지 캡션 구조의 textual decoder의 발전 동향에 대해 다루고자 한다. 캡션을 생성하기 위해 LSTM 변형 네트워크를 적용한 것부터 병렬 처리를 위해 사용한 CNN과 Transformer 그리고 사전 학습 언어 모델인 BERT와 GPT를 사용하기 까지의 발전 과정과 각 방법의 장단점을 소개한다.

가. LSTM 기반의 모델

언어는 순차적인 구조를 가지므로 RNN 계열의 네트워크는 자연스럽게 문장 생성을 처리하는 데 적합하다. 특히 LSTM 변종 네트워크는 초기 이미지 캡션 디코더에서 표준으로 사용되었다. Vinyals et al. [13]은 시각적 인코딩이 LSTM의 초기 은닉 상태로 사용되어 출력 캡션을 생성한다. 각 시간 단계에서, 은닉 상태를 vocabulary와 동일한 크기의 벡터로 투영하는 것에 대해 softmax 활성화 함수를 적용하여 단어를 예측한다. 훈련 중, 입력 단어는 ground-truth 문장에서 가져오고 추론 단계에서는 입력 단어는 이전 단계에서 생성된 것이다. 얼마 지나지 않아 Xu et al. [23]는 additive attention mechanism을 도입했다. 이는 static global 벡터를 대체하고 단어와 시각적 contents 간의 정렬을 개선하는 이미지의 동적 및 time-varying 표현이다. 이 경우 이전의 은닉 상태는 시각적 features X에 대한 attention 메커니즘을 적용하여 출력 단어 예측을 담당하는 MLP에 공급되는 context vector를 계산한다. LSTM은 multi-



(그림 5) LSTM 구조 [13]



(그림 6) Aneya et al. [29] 제안한 모델의 아키텍처

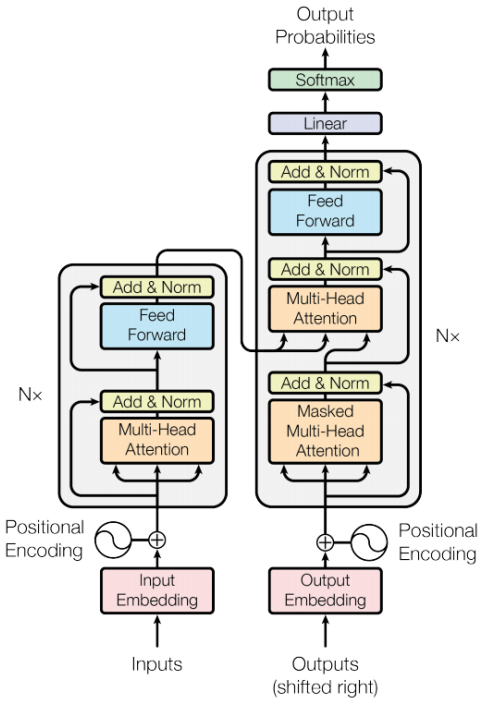
layer 구조로 확장하여 higher-order 관계를 캡처하는 능력을 증가할 수 있다. Donahue et al.[16]는 처음으로 2 계층 LSTM을 캡션용 언어 모델로 제안했으며, 첫 번째 계층의 은닉 상태가 두 번째 계층이 입력이 되도록 두 계층을 쌓았다. Huang et al. [20]은 시각적 self-attention 위에 또 다른 단계를 계산하는 Attention on Attention 연산자로 LSTM을 추가했다.

나. Convolution 기반의 모델

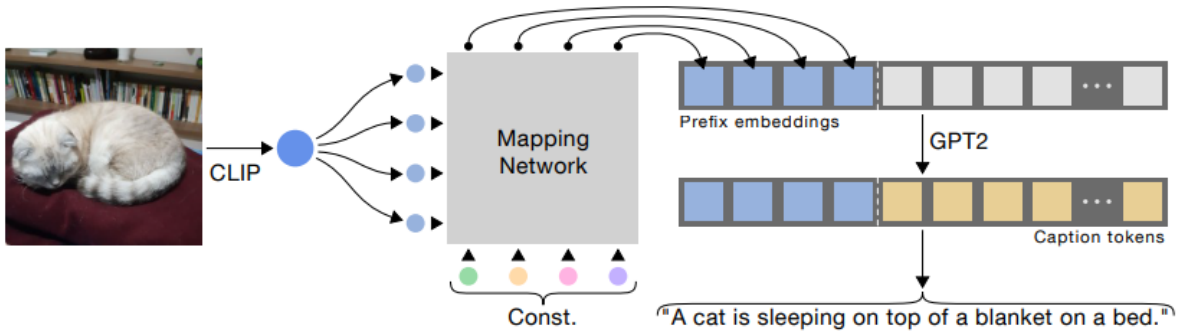
기존의 LSTM 네트워크를 사용한 경우 모델이 우수한 성능을 보였지만 LSTM 네트워크는 복잡하고 순차적이라는 본질적인 특성이 존재한다. Aneya et al. [29]은 LSTM의 단점을 극복하기 위해 RNN 계열의 순차적인 특성을 가진 네트워크를 사용하는 대신 CNN을 언어 모델로 사용하는 방법을 제안했다. 추가적으로 attention 메커니즘을 사용해 공간적 이미지 feature를 활용했다.

다. Transformer 기반의 모델

Convolution 네트워크의 병렬 훈련이라는 장점에도 불구하고 Transformer 아키텍처의 등장으로 인해 언어 모델에서 Convolution 연산자의 사용은 큰 인기를 얻지 못했다. "Attention is all you need" 논문에서 Vaswani et al.[10]이 제안한 fully-attentive 패러다임은 언어 생성의 관점을 완전히 바꾸어 놓았다. 원래의 Transformer 디코더는 수정 없이 이미지 캡션 모델에 사용되었다 [19].



(그림 7) Transformer 구조 [10]



(그림 8) Mokady et at. [28]이 제안한 모델의 아키텍처

라. 사전 학습 언어 모델 기반의 모델

얼마 지나지 않아 Transformer 모델은 BERT 및 와 GPT와 같은 NLP에서의 혁신적은 모델의 building block으로 사용되었고, 많은 언어 이해 작업을 위한 사실상의 표준 아키텍처가 되었다. Li et al. [22]는 이미지와 텍스트 간의 의미 정렬을 용이하게 하기 위해 객체 태그를 anchors points로 포함하는 BERT 유사 아키텍처인 Oscar를 제안했다. 이미지에서 탐지된 객체 태그를 시각-언어 결합 표현에서 더 나은 정렬을 학습하기 위한 anchors points로 사용하는 방법을 도입했다. 이를 위해, 그들의 모델은 입력 이미지-텍스트 쌍을 단어 토큰-객체 태그-영역 feature 3가지로 표현하며 여기서 객체 태그는 object detector가 제안하는 텍스트 클래스이다. Zhou et al. [21]는 시각적 및 텍스트 양식을 이미지 캡션을 위한 BERT와 유사한 아키텍처로 융합하는 통합 모델을 개발했다. 이 모델은 인코딩과 디코딩 모두에 대해 shared multi-layer 트랜스포머 인코더 네트워크로 구성되며, 이미지 캡션 쌍의 대규모 말뭉치에서 사전 훈련된 다음 토큰 시퀀스를 오른쪽 마스킹하여 단방향 생성 프로세스를 시뮬레이션하여 이미지 캡션을 위해 미세 조정된다. Mokady et al.[28]은 시각적 인코더로 CLIP 모델의 시각적 인코더를 사용하고 디코더로 GPT2를 사용했다.

III. 이미지 캡셔닝 기술 성능 비교

II장에서는 이미지 캡션 모델의 다양한 visual encoder와 textual decoder의 동향에 대해 살펴보았다. 각 모델과 방법은 이미지 feature와 캡션 생성 그리고 학습 속도 등의 측면에서 장단점을 가지고 있는데 Stefanini et al. [30]은 이를 동일한 조건에서의 실험을 통해 비교하였다. (표 1)를 통해 위의 실험 결과를 훈련 시간과 성능 지표에 따라 확인할 수 있다. 현재까지의 연구를 살펴보면 표준 평가 지표에서 VinVL 모델이 가장 뛰어난 성능

		Standard Metrics						Diversity Metrics				Embedding-based Metrics			Learning-based Metrics			
	#Params (M)	B-1	B-4	M	R	C	S	Div-1	Div-2	Vocab	%Novel	WMD	Alignment	Coverage	TIGEr	BERT-S	CLIP-S	CLIP-S^{Ref}
Show and Tell† [23]	13.6	72.4	31.4	25.0	53.1	97.2	18.1	0.014	0.045	635	36.1	16.5	0.199	71.7	71.8	93.4	0.697	0.762
SCST (FC)‡ [38]	13.4	74.7	31.7	25.2	54.0	104.5	18.4	0.008	0.023	376	60.7	16.8	0.218	74.7	71.9	89.0	0.691	0.758
Show, Attend and Tell† [42]	18.1	74.1	33.4	26.2	54.6	104.6	19.3	0.017	0.060	771	47.0	17.6	0.209	72.1	73.2	93.6	0.710	0.773
SCST (Att2in)‡ [38]	14.5	78.0	35.3	27.1	56.7	117.4	20.5	0.010	0.031	445	64.9	18.5	0.238	76.0	73.9	88.9	0.712	0.779
Up-Down‡ [58]	52.1	79.4	36.7	27.9	57.6	122.7	21.5	0.012	0.044	577	67.6	19.1	0.248	76.7	74.6	88.8	0.723	0.787
SGAE [71]	125.7	81.0	39.0	28.4	58.9	129.1	22.2	0.014	0.054	647	71.4	20.0	0.255	76.9	74.6	94.1	0.734	0.796
MT [72]	63.2	80.8	38.9	28.8	58.7	129.6	22.3	0.011	0.048	530	70.4	20.2	0.253	77.0	74.8	88.8	0.726	0.791
AoANet [79]	87.4	80.2	38.9	29.2	58.8	129.8	22.4	0.016	0.062	740	69.3	20.0	0.254	77.3	75.1	94.3	0.737	0.797
X-LAN [80]	75.2	80.8	39.5	29.5	59.2	132.0	23.4	0.018	0.078	858	73.9	20.6	0.261	77.9	75.4	94.3	0.746	0.803
DPA [83]	111.8	80.3	40.5	29.6	59.2	133.4	23.3	0.019	0.079	937	65.9	20.5	0.261	77.3	75.0	94.3	0.738	0.802
AutoCaption [107]	-	81.5	40.2	29.9	59.5	135.8	23.8	0.022	0.096	1064	75.8	20.9	0.262	77.7	75.4	94.3	0.752	0.808
ORT [77]	54.9	80.5	38.6	28.7	58.4	128.3	22.6	0.021	0.072	1002	73.8	19.8	0.255	76.9	75.1	94.1	0.736	0.796
CPTR [92]	138.5	81.7	40.0	29.1	59.4	129.4	-	0.014	0.068	667	75.6	20.2	0.261	77.0	74.8	94.3	0.745	0.802
$\mathcal{M}^2$ Transformer [81]	38.4	80.8	39.1	29.2	58.6	131.2	22.6	0.017	0.079	847	**78.9**	20.3	0.256	76.0	75.3	93.7	0.734	0.792
X-Transformer [80]	137.5	80.9	39.7	29.5	59.1	132.8	23.4	0.018	0.081	878	74.3	20.6	0.257	77.7	75.5	94.3	0.747	0.803
Unified VLP [101]	138.2	80.9	39.5	29.3	59.6	129.3	23.2	0.019	0.081	898	74.1	**26.6**	0.258	77.1	75.1	**94.4**	0.750	0.807
VinVL [103]	369.6	**82.0**	**41.0**	**31.1**	**60.9**	**140.9**	**25.2**	**0.023**	**0.099**	**1125**	77.9	20.5	**0.265**	**79.6**	**75.7**	88.5	**0.766**	**0.820**

(표 1) 성능 비교 [30]

을 보여준다.

IV. 결론

이미지 캡션은 컴퓨터 비전(Computer Vision)과 자연어 생성(Natural Language Generation)의 어려움을 통합하기 때문에 machine intelligence에서 본질적으로 복잡한 작업이다. 대부분의 접근 방식은 시각적 인코딩 및 언어 모델링 단계를 구별하는 반면, BERT와 유사한 아키텍처의 singe-stream 트렌드에는 시각적 및 텍스트 데이터의 early-fusion이 수행된다. 이러한 전략을 사용하면 놀라운 성능을 달성할 수 있지만 일반적으로 대규모 사전 훈련과 필요하게 된다. 따라서 사전 훈련을 통한 표준 인코더-디코더 방법이 유사한 결과를 얻을 수 있는지 조사할 가치가 있다. 그럼에도 불구하고, 모델 설계자와 최종 사용자 모두에게 고전적인 two-stream 패러다임에 기반한 방법이 더 설명하기 쉽다. 제시된 관련 연구 및 실험 비교는 지난 몇 년 동안의 성능 향상을 보여준다. 그러나 정확성, 견고성 및 일반화 결과가 만족스럽지 못하기 때문에 아직 해결해야 할 과제가 많이 남아 있다. 마찬가지로, 캡션의 정확성과 자연스러움 및 다양성의 요구 사항은 아직 충족되지 않았다. 이와 관련하여 이미지 캡션은 사람과 기계 간의 상호 작용을 개선하기 위해 고안되었기 때문에 사용자를 루프에 포함할 가능성이 있다. 제시된 동향 분석을 바탕으로 이미지 캡션 분야에 대한 발전 방향을 추적할 수 있다.

참고문헌

[1] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification

with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105.

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

[4] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

[5] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *arXiv preprint arXiv:1602.07332* (2016).

[6] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[7] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *European conference on computer vision*. Springer, Cham, 2014.

[8] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *arXiv preprint arXiv:2103.00020* (2021).

[9] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[10] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[11] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[12] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

[13] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[14] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[15] Mao, Junhua, et al. "Deep captioning with multimodal recurrent neural networks (m-rnn)." *arXiv preprint arXiv:1412.6632* (2014).

[16] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[17] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015): 91-99.

[18] Ren, Shaoqing, et al. "Faster R-CNN: towards real-time object detection with region proposal networks." *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016): 1137-1149.

[19] Herdade, Simao, et al. "Image captioning: Transforming objects into words." *arXiv preprint arXiv:1906.05963* (2019).

[20] Huang, Lun, et al. "Attention on attention for image captioning." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[21] Zhou, Luowei, et al. "Unified vision-language pre-training for image captioning and vqa." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 07. 2020.

[22] Li, Xiujun, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." *European Conference on Computer Vision*. Springer, Cham, 2020.

[23] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. PMLR, 2015.

[24] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[25] Yang, Xu, Hanwang Zhang, and Jianfei Cai. "Learning to collocate neural modules for image captioning." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[26] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[27] Liu, Wei, et al. "Cptr: Full transformer

network for image captioning." *arXiv preprint arXiv:2101.10804* (2021).

[28] Mokady, Ron, Amir Hertz, and Amit H. Bermano. "ClipCap: CLIP Prefix for Image Captioning." *arXiv preprint arXiv:2111.09734* (2021).

[29] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[30] Stefanini, Matteo, et al. "From show to tell: A survey on image captioning." *arXiv preprint arXiv:2107.06912* (2021).