객체 검출 기반 다중 모달리티 어텐션 향상 *

김용일 O1 , 윤현구 1, 황예린 2, 정교민 1,2

1 서울대학교 전기정보공학부, 2 서울대학교 협동과정 인공지능

{miles94, youraredead, dpfls589, kjung}@snu.ac.kr

Improving cross-modal attention via object detection

Yongil Kim O1, Hyeongu Yun1, Yerin Hwang2, Kyomin Jung1,2

1 Department of Electrical and Computer Engineering, Seoul National University

2 Interdisciplinary Program in Artificial Intelligence, Seoul National University

요 약

최근 이미지 캡셔닝과 같이 서로 다른 모달리티를 하나의 모델에서 다루는 멀티 모달리티에 대한 연구가 활발히 진행되고 있다. 이러한 멀티 모달리티 연구에서는 서로 다른 두 모달리티에 대한 정보를 혼합하기 위해 cross-modal attention 이 많이 사용된다. 하지만 현존하는 대부분의 모델들은 cross-modal attention을 차용하여 종단간 학습을 통한 간접적인 학습에만 의존하고, 직접적인 어텐션에 대한 개선은 하지 않는다. 본 연구에서는, 멀티 모달리티 중 특히 이미지 정보와 텍스트 정보를 다루는 Vision-and-Language 태스크에 객체 검출 모델을 차용하여 cross-modal attention 을 직접적으로 향상시키는 새로운 방법론을 제시한다. 본 논문에서 제시하는 방법론은 다양한 데이터셋에 대한 실험의 모든 성능 지표에서 기존의 베이스라인의 성능을 크게 향상시키며, 추후 Vision-and-Language 의 어느 태스크에도 적용 가능한 확장성이 높은 방법론이다.

1. 서 론

최근 영상 정보와 텍스트 정보, 음성 정보 등 다른 특성을 갖는 여러 모달리티 (Modality) 를 다루는 멀티 모달리티 (Multi-modality) 모델에 관한 연구가 활발하다. 특히 Vision-and-Language 라고 부르는 연구 분야는 이미지 정보와 텍스트 정보를 입력으로 받아 원하는 목적을 이루는 연구로, 이미지 질의 응답 (VisualQA), 이미지 캡셔닝 (Image Captioning) 등의 태스크를 포함한다. 다른 멀티 모달리티 연구와 마찬가지로, 이미지와 텍스트의 서로 다른 특성을 갖는 두 모달리티를 함께 다루는 연구이기 때문에, 두 모달리티를 잘 혼합하여 원하는 출력을 내놓는 모델을 설계해야 한다.

최근 가장 좋은 성능을 보이고 있는 멀티 모달리티 모델들은 두 가지 이상의 모달리티를 혼합하기 위하여 대부분 어텐션 기법을 활용한다. 어텐션 기법은 모델로 하여금 가장 중요한 부분을 스스로 학습하게 도와주며, 두 모달리티 간에 서로 중요한 부분을 공유할 수 있게 도와주는 역할을 한다. 이렇게 두 모달리티 사이에 적용되는 어텐션이 cross-modal attention이다. 특히, 최근 어텐션 기법을 기반으로 한 트랜스포머 (transformer) 모델들이 멀티 모달리티 연구에서도 강세를 보이고 있기 때문에, 더욱 cross-modal attention 은 멀티 모달리티 모델에서 필수적인 요소가 된다. 또한, Vision-and-Language 연구에서 역시 cross-modal attention 을 활용하여 결과를 얻어낸 연구가 많다[1][2].

하지만 현존하는 대부분의 Vision-and-Language 모델들에서는 cross-modal attention 모듈을 단순히 차용하기만 할 뿐, 그것을 개선하려는 시도가 존재하지 않는다. 즉, 기존에는 어텐션에 대하여 목적 함수를 이용한 직접적인 학습을 한 연구는 많이 이뤄지지 않았다. 본 연구에서는 Vision-and-language 연구에서 cross-modal attention 을 직접적으로 개선하여, 모델 전체의 성능을 끌어올리는 연구를 제안한다.

구체적으로, 본 연구에서는 객체 검출 모델 (Object Detection model)을 이용하여 어텐션을 개선하는 연구를 제안한다. 최근 detectron2 등 사람의 능력을 상회하거나 거의 유사한 성능을 보이는 강력한 객체 검출 모델들이 등장하였다. 우리 모델은 이렇게 강력한 성능의 객체 검출 모델을 차용하여, pseudo self-supervised learning 을 이용하여 cross-modal attention을 개선하는 방법론을 제안한다.

본 연구에서 제안한 방법론을 이미지 캡셔닝 태스크에 적용하여 여러 데이터셋에서 실험을 수행한 결과, 모든 데이터셋과 모든 성능 지표에서 기존 베이스라인 모델에 비해 획기적인 성능 향상이 이루어졌다. 본 연구의 방법론은, cross-modal attention 의 성능을 개선하고자하는 새로운 모델 학습 방법론(training scheme)이기 때문에, 모든 Vision-and-Language Task 에 Model-agnostic 하게 적용될 수 있고, 따라서 본 연구진은 제안된 방법론을 다양한 분야에 적용하는 추가 연구를 진행할 계획이다.

* 이 연구는 서울대학교 자동화연구소와 2022년도 BK21 FOUR 정보기술 미래인재 교육연구단, 그리고 삼성전자의 지원을 받아 수행된 연구결과임.

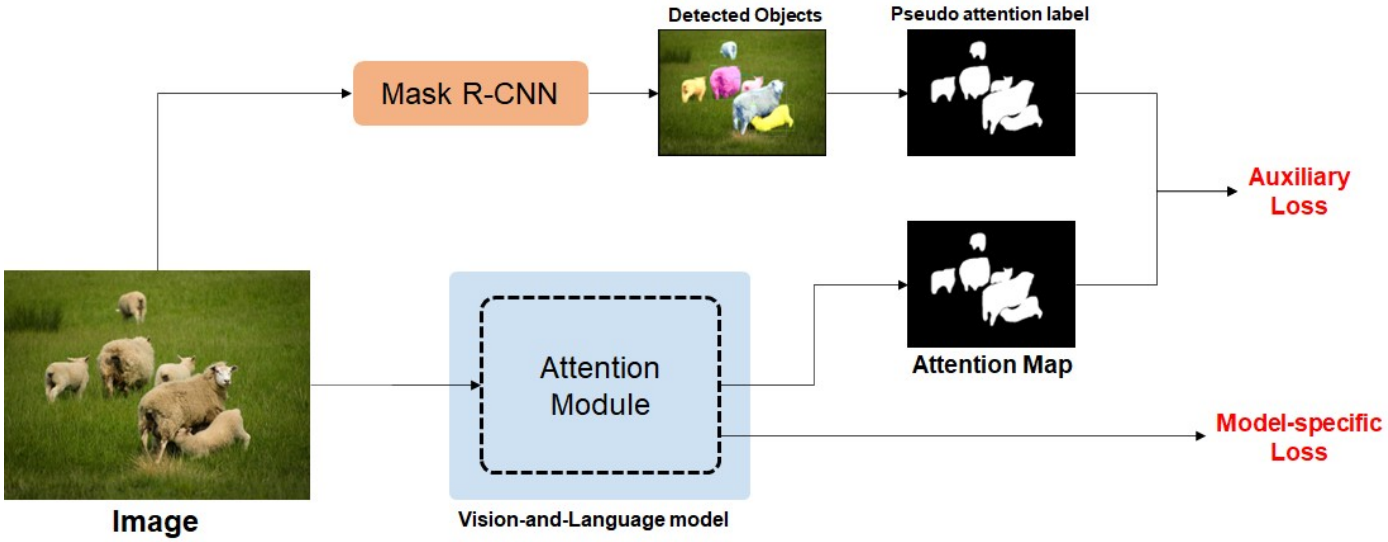

그림 1. 본 연구에서 제안하는 모델

2. 관련 연구

2.1 Cross-modal Attention 의 개선

멀티 모달리티 연구에서 cross-modal attention 의 효과를 향상시키기 위한 여러 연구가 존재한다. 한 연구[3]에서는 두 모달리티 사이의 특성 차이를 줄여 두 모달리티 사이의 어텐션의 효과를 향상시키려는 연구가 존재한다.

그러나 어텐션에 대한 직접적인 개선에 관한 연구가 많이 이뤄지지 않고 있다. 현존하는 대부분의 모델에서 어텐션 기법은 종단 간 학습 (End2End Learning) 으로 학습된다는 점과 그에 따라 직접적인 목적 함수 설계가 쉽지 않다는 점이 그 이유다. 본 연구에서는 객체 검출 모델을 활용한 목적 학수를 설계를 통하여 cross-modal attention 으로의 직접적인 개선 방법을 제안한다.

2.2 Surrogate IOU

IOU(Intersection over Union) 메트릭은 객체 검출 모델이 출력하는 결과와 실제 라벨 사이의 겹치는 정도를 통하여 검출 성능을 나타내는 지표로 활용된다. IOU 메트릭에 대한 성능을 높이기 위하여 이를 직접적인 목적 함수로 사용하고자하는 시도가 있었지만, IOU 메트릭은 미분 불가능(Indifferentiable)하기 때문에, 대부분의 Gradient-based 모델에서는 사용이 불가능하다. 이에 여러 연구[4][5]에서 이 메트릭을 미분 가능한 대체 함수 (surrogate function)를 통해 직접적으로 학습에 활용하였다.

본 연구에서는 Vision-and-Language 모델의 cross-modal attention 에 surrogate IOU 함수를 이용한다. 기존 연구들이 객체 검출 모델 자체의 성능 향상에 초점을 맞추었다면, 본 연구에서는 어텐션 모듈이 내놓는 결과와 객체 검출 모델이 내놓는 결과 사이의 목적함수 설계에 surrogate IOU 가 활용된다는 점이 다르다.

3. 제안된 모델

본 연구에서 제안하는 cross-modal attention 향상 방법론은 모든 Vision-and-Language 연구에 적용될 수 있다. 본 연구에서 제안하는 모델은 그림 1에서 볼 수 있다. 객체 검출 모델로는 Mask R-CNN[6]을 차용한다. Mask R-CNN 은 이미지를 입력으로 받아 검출되는 객체의 클래스 라벨과 Mask 를 출력으로 내놓게 된다. 본 연구에서는 이렇게 자연어로 주어지는 클래스 라벨을 Vision-and-Language 태스크의 입력 텍스트와 비교한다. 예를 들어, 이미지 캡셔닝의 경우 Mask R-CNN이 검출한 객체가 주어진 캡션에 들어있다면, 그것을 critical object로써 Mask 를 추출한다. 이것을 Pseudo Attention Label 로써 활용한다.

한편, 그림 1의 아래 경로처럼, 이미지가 모델의 Attention Module 을 통해 각 워드 토큰에 해당하는 Attention Map 을 출력으로 내어놓는다. 이후 입력으로 주어지는 텍스트의 워드 토큰과 Mask R-CNN이 예측한 클래스 라벨에 모두 해당하는 Critical object 워드 토큰에 해당하는 Attention Map 을 추출한다. 이 Attention Map 과 앞서 Mask R-CNN 이 출력한 Pseudo Attention Label 을 이용하여 Surrogate IOU 로 Auxiliary Loss를 구성한다. Auxiliary Loss 의 수식은 아래와 같다:

$$\mathcal{L}_{Aux} = \overline{\Delta_{J_1}}(m(F))$$

F는 attention map feature 값이고, m 은 hinge Loss 이며, $\overline{\Delta_{J_1}}$ 은 Jaccard Loss 의 Lovasz extension 이다. 따라서, Auxiliary Loss 는 Jaccard Loss의 Lovasz hinge extension 이다. 이후 이 Loss를 원래의 모델의 목적 함수에 추가적으로 반영한다. 원래의 목적함수인 Model-specific Loss 는 Vision-and-Language 의 태스크마다 달라진다. 본 연구에서 제안하는 방법은 모델에 상관없이 적용될 수 있는 Model-agnostic 방법론이기 때문에, 추가적인 Auxiliary Loss 를 통해 어떤 Loss 에도 적용될 수 있다.

$$\mathcal{L} = \mathcal{L}_{model-specific} + \lambda_{Aux}\mathcal{L}_{Aux}$$

본 연구에서는 λ_{Aux} 은 0.1 로 설정하였다.

표 1. 이미지 캡셔닝 실험 결과

Dataset	Model	BLEU-1	BLEU-4	CIDEr	METEOR	ROUGE-L	SPICE
Flickr8k	*Show-Attend-Tell	0.6460	0.2339	0.5874	0.2032	0.5306	0.1495
	+ ICA Loss	**0.6576**	**0.2453**	**0.5939**	**0.2256**	**0.5479**	**0.1713**
Flickr30k	*Show-Attend-Tell	0.6557	0.2334	0.4789	0.1930	0.5085	0.1330
	+ ICA Loss	**0.6597**	**0.2351**	**0.4798**	**0.1963**	**0.5132**	**0.1351**
MSCOCO	*Show-Attend-Tell	0.7050	0.3036	0.9378	0.2472	0.5684	0.1776
	+ ICA Loss	**0.7110**	**0.3149**	**0.9392**	**0.2488**	**0.5792**	**0.1793**

*our Implementation

4. 이미지 캡셔닝 실험과 결과

본 연구에서 진행한 Vision-and-Language Task 로는 이미지 캡셔닝을 선택하여 실험을 진행하였다. Baseline 모델로는 Show-attend-tell[7] 모델을 활용한다. 이 베이스라인 모델은 어텐션 모듈을 활용하여 캡션의 각 워드 토큰에 해당하는 이미지와의 어텐션을 활용하는 모델이다. 실험에 사용된 데이터셋은 이미지 캡셔닝에서 가장 많이 사용되는 Flickr8k, Flickr30k, MSCOCO 데이터셋을 활용하고, 성능 지표로는 자연어 생성 성능 지표인 BLEU-1, BLEU-4, CIDEr, METEOR, ROUGE-L, SPICE 를 사용한다. 기존 베이스라인 모델과 본 연구에서 제안한 방식의 모델 모두 직접 구현한 모델을 활용하며, 성능 지표는 PyCoCo에 구현된 공식적인 성능지표를 활용한다. 실험 시 Mask R-CNN이 발견한 클래스 라벨 중 캡션에 속하는 Critical Object 는 인스턴스당 평균적으로 2.3개에 해당하였고, 해당 Critical object 들의 Loss를 mean average 를 취하여 Loss 를 구성하였다.

실험 결과는 표 1에서 볼 수 있다. ICA Loss 는 Improving Cross-modal Attention Loss 로 본 연구에서 추가한 Auxiliary Loss 이다. 본 연구에서 제안하는 Auxiliary Loss 를 추가한 모델이 모든 데이터셋과 모든 성능지표에서 훨씬 좋은 성능을 보이는 것을 확인할 수 있다. 특히 Flickr8K 와 MSCOCO의 데이터셋에서 ICA Loss 를 추가하였을 때 BLEU-4 와 ROUGE-L 등의 성능지표에서 획기적인 성능 향상을 보인다. 이를 통해 본 연구에서 제안하는 어텐션의 직접적인 개선이 모델 성능 향상으로 직접적으로 긍정적인 영향을 준 것을 확인할 수 있다.

5. Broader Impact

앞서 언급했듯이, 본 연구에서 제안하는 방식은 모든 Vision-and-Language 연구에 Model-agnostic 한 방식으로 적용될 수 있다. 객체 검출 모델이 출력하는 Mask 와 그에 해당하는 자연어로 생성된 클래스 라벨을 Vision-and-Language 태스크에서 활용될 수 있다. 예를 들어, 본 연구에서 활용한 이미지 캡셔닝 외에도, 이미지 질의 응답(Visual Question Answering) 에서는 질문 속의 Critical object 를 찾아내는 용도와 그에 해당하는 cross-modal attention 의 직접적인 향상을 도모할 수 있고, 자연어 처리를 활용한 이미지 추론(Natural Language for Visual Reasoning) 에서는 두 개의 이미지에 대한 객체 검출 모델의 예측 클래스가 겹치는 객체가 자연어에 있을 경우, 중요한 역할을 할 수 있다고 판단할 수 있다. 그 외에도 텍스트 모달리티와의 연계가 필요한 모든 멀티 모달리티 태스크에 적용이 가능하여, 본 연구가 추후 확장 측면에서 그 가치가 굉장히 높다.

6. 결론

본 논문에서는 기존의 Vision-and-Language 연구에 객체 탐지 모델을 차용하여 성능을 높이는 방법을 제안하였다. 객체 탐지 모델의 출력 마스크와 모델의 어텐션맵에 미분 가능한 Surrogate IOU 목적함수를 적용하여, 직접적인 학습을 통한 어텐션 개선 방법론을 제시하였다. 본 연구에서 제안하는 방법은 추후 여러 태스크로의 확장가능성을 지닌 연구이며, 이미지 캡셔닝 태스크에 대하여 실험적으로 큰 성능 향상을 보여 그 우수성을 검증하였다.

7. 관련 논문

[1] Wei, Xi, et al. "Multi-modality cross attention network for image and sentence matching." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[2] Ye, Linwei, et al. "Cross-modal self-attention network for referring image segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[3] 김용일, et al. "모달리티 정렬을 통한 비디오 질의 응답 시스템 개선." *한국정보과학회 학술발표논문집* (2021): 570-572.

[4] Berman, Maxim, Amal Rannen Triki, and Matthew B. Blaschko. "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[5] Nagendar, Gattigorla, et al. "Neuro-IoU: Learning a Surrogate Loss for Semantic Segmentation." BMVC. 2018.

[6] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

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