Transformer 를 이용한 Image Captioning

김동훈, Xu Jiajie, 심병효*
서울대학교

dhkim@islab.snu.ac.kr, jiajiexu@islab.snu.ac.kr, *bshim@islab.snu.ac.kr

A Study on the image captioning using transformer

Donghoon Kim, Jiajie Xu, Byonghyo Shim*
Seoul National University

요 약

본 논문은 vision 과 language domain 의 정보를 동시에 학습하는 image captioning task 에 대해서 Transformer 를 적용하는 기법에 대해서 다룬다. Vision 과 language domain 을 동시에 학습하는 multi-modality 학습 기법은 image captioning task 뿐만 아니라 text-to-image retrieval, video question and answering, video moment retrieval 등 다양한 영역에서 사용된다. 그런데 수많은 task 들 중에서도 image captioning task 는 이러한 수많은 multi-modality 중에서도 두가지 도메인에 신경망의 이해도를 측정하는 데에 있어 가장 많은 주목을 받고 있다. 이는 현 시점의 많은 연구들이 image captioning task 를 pretrain task 를 이용하고 down stream task 로 다양한 task 를 가능하게 하고 있다는 점에서 확인 가능하다. 본 논문은 이러한 image captioning task 에 대해서 transformer 구조를 활용하여 coco dataset 에 대해서 scratch 로부터 end-to-end 로 학습시켜 image captioning task 를 수행 해 보았다.

Ⅰ. 서 론

Vision domain 과 language domain 은 기계 학습 분야에 있어서 가장 큰 축들 중에 하나이다. Vision 도메인의 경우 CNN 모델을 통한 image classification, detection 그리고 segmentation 과 같은 다양한 ask 에서의 괄목할 만한 성과를 이루어 왔다. Language domain 역시 최근 Transformer 를 이용한 BERT[1], GPT[2], 그리고 다양한 파생 모델들이 탄생하고 translation, text generation, text sentimental classification, grammatical error classification, 그리고 question-answering 등의 task 들에서 human level 을 뛰어 넘는 놀라운 성과를 보이고 있다. 특히 OpenAI 를 통해서 발표된 GPT3[3]와 같은 초대형 모델이 앞에서 언급된 수많은 task 들에 대해서 zero-shot 성능이 State-of-the-art(SOTA)를 달성하는 것을 보여주며 natural language domain 에서 기계학습의 무한한 가능성을 보여주고 있다.

Vision 과 language domain 각각에서의 놀라운 성과에도 불구하고 두 domain 을 통합시키는 multi-modality learning 은 그만한 관심을 받지 못했다. Multi-modality learning 이란 인간의 다인지 학습 능력을 모방하여 다양한 서로 다른 domain 의 신호를 통합하여 처리하는 학습 방법을 의미한다. video question answering 이나 image captioning 과 같이 두가지 서로 다른 domain 의 data 를 모두 사용하는 task 들의 경우 두 domain 의 통합이 반드시 필요하다. 뿐만 아니라 Multi-modality learning 은 각 domain 의 단점을 보완해 주고 장점을 부각시키면서 성능을 향상시킨다. 예를 들어 image classification 에서 language representation 을 이용하여 학습을 시킴으로써 zero-shot capability 를 향상시키는[4] 등의 domain specific 한 task 에서의 다양한 문헌이 발표되고 있다.

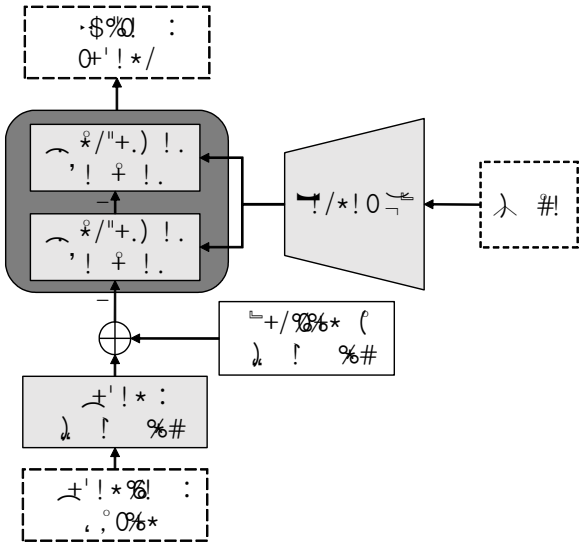

그림 1 모델 구조

이렇게 multi-modality 의 중요성이 부각되고 있는 가운데 해당 분야에서 가장 활발히 연구되고 있는 분야는 vision 과 language 의 통합 학습에 대한 분야이다. Vision 과 language 통합의 대표적인 task 들에는 vision 과 language domain 을 통합하여 수행하는 image captioning, text-to-image retrieval, video question and answering, video moment retrieval 등이 있다. 이러한 수많은 task 들 중 최근 발표되는

연구에서는 language model 에서 영감을 받아 image caption generation[6]과 masked token prediction [6, 8]을 pretrain 으로 활용하고, 나머지 task 들을 downstream task 로 fine-tuning 한다.

본 연구에서는 vision-and-language learning 에서 널리 pretrain task 로 이용되고 있는 image captioning 을 위해 transformer 와 resnet-50 을 scratch 부터 end-to-end 로 학습시켜 image captioning task 를 수행했다.

Ⅱ. 본론

본 연구에서 사용한 모델의 구조는 그림 1 과 같다. Transformer decoder 의 경우 original transformer[9]의 decoder 를 그대로 사용하였다. Resnet-50 도 마찬가지로 기존 model 의 구조를 그대로 사용했다. Tokenizer 의 경우 Byte Pair Encoding (BPE)를 통해 COCO caption 에서 추출한 10K 사이즈의 corpus 를 생성하여 이용하였다.

우선 caption 과 image 입력을 받으면, caption 은 tokenizer 로 token 들로 변환을 한다. 이때 캡션 token 맨 앞에 [sos] 토큰을 붙인다. 그리고 hidden size 가 1024 인 token embedding 으로 변환하고 같은 사이즈의 learnable parameter 인 positional embedding 을 더해주어 Transformer Decoder 에 Query 로 입력한다. Image 의 경우 높이와 너비를 모두 224 로 Crop 한다. 그리고 Resnet-50 을 통과해서 생성된 7*7*2048 (H*W*C)의 feature 를 channel 크기가 1024 가 되도록 Linear projection 시켜준 뒤, 49 개의 feature pixel 데이터들을 Transformer Decoder 의 Key 및 Value 로 입력한다. Output 은 shifted token 을 label 로 하여 Cross-entropy 를 이용하여 학습한다.

batch size 의 경우 64 로 하고 270 epoch 동안 학습시켰다. 학습에 사용된 optimizer 는 Stochastic Gradient Decent (SGD) 이다. momentum 은 0.9 를 weight decay 는 0.0001 이다. Learning rates 은 Resnet 의 경우 0.2 를 Transformer 의 경우에는 0.001 을 사용했다. 둘다 10 epoch 동안을 warm up step 으로 하는 cosine learning rate decay 를 이용했다.

이렇게 학습된 모델의 metric 은 CIDER score 와 SPICE score 를 통해 확인했는데, 결과는 표 1 과 같다.

Metric name	Score
CIDER	93.7
SPICE	18.0

표 1 Image captioning metric

Ⅲ. 결론

본논문에서는 Transformer 를 이용하여 scratch 부터 end-to-end 로 학습시킨 image captioning 을 확인 해 보았다. 입력 이미지에 대한 출력 caption 은 그림 2 와 같이 확인 가능하며 표 1 에서 나온 성능을 통해 사람이 labeling 한 caption 과 구분이 거의 불가능 함을 알 수 있다. 추후에는 model 구조를 변경하고 image 단에서의 loss 를 추가하여 학습을 시켜 성능을 향상시켜 볼 예정이다.

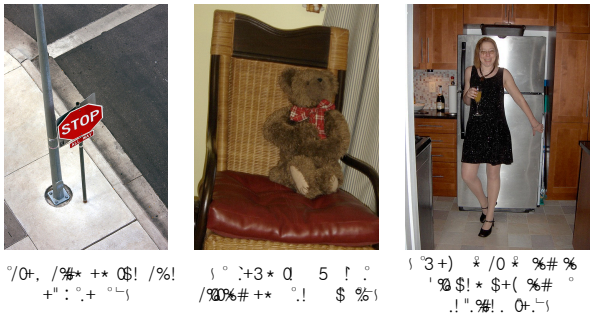

그림 2 생성된 캡션 예시

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT 연구센터육성지원사업의 연구결과로 수행되었음. (IITP-2021-2017-0-01637)

참 고 문 헌

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

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

[3] Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

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

[5] Liu, Yang, et al. "Use what you have: Video retrieval using representations from collaborative experts." arXiv preprint arXiv:1907.13487 (2019).

[6] Lei, Jie, et al. "Less is more: Clipbert for video-and-language learning via sparse sampling." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[7] Desai, Karan, and Justin Johnson. "Virtex: Learning visual representations from textual annotations." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

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

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