Transformer 기반의 Clustering CoaT 모델 설계

방지현[1], 박준[1], 정세훈[2], 심춘보[1]
[1]순천대학교 IT-Bio 융합시스템전공
[2]안동대학교 창의융합학부
jihyeon2974@naver.com, todnehfdl@naver.com jungsh@anu.ac.kr, cbsim@scnu.ac.kr

Design of Clustering CoaT Vision Model Based on Transformer

Ji-Hyeon Bang[1], Se-Hoon Jung[2], Chun-Bo Sim[1]
[1]Interdisciplinary Program in IT-Bio Convergence System, Sunchon National University, Sunchon National University
[2]Dept. of Creative Convergence, Andong National University

요 약

최근 컴퓨터 비전 분야에서 Transformer를 도입한 연구가 활발히 연구되고 있다. 이 모델들은 Transformer의 구조를 거의 그대로 사용하기 때문에 확장성이 좋으며 large 스케일 학습에서 매우 우수한 성능을 보여주었다. 하지만 Transformer를 적용한 비전 모델은 inductive bias의 부족으로 학습 시 많은 데이터와 시간을 필요로 하였다. 그로 인하여 현재 많은 Vision Transformer 개선 모델들이 연구되고 있다. 본 논문에서도 Vision Transformer의 문제점을 개선한 Clustering CoaT 모델을 제안한다.

1. 서론

최근 컴퓨터 비전 분야에서 Natural Language Processing(NLP) 분야의 Transformer를 도입한 연구가 활발히 연구되고 있다. 처음 Attention 기법은 컴퓨터 비전 분야에서 Convolutional Neural Networks(CNN)과 함께 사용되거나 CNN의 부수적인 요소로 사용되었다. 하지만 Vision Transformer 모델의 등장 이후, 비전 모델은 CNN 대신 Transformer를 적용하는 모델들이 대거 등장하였다. 이 모델들은 Transformer의 구조를 거의 그대로 사용하기 때문에 확장성이 좋으며 large 스케일 학습에서 매우 우수한 성능을 보여주었다. 하지만 Transformer를 적용한 비전 모델은 inductive bias의 부족으로 인해 CNN보다 많은 데이터를 요구하여 학습 시 많은 시간이 필요하다[1]. 이와 같은 Vision Transformer의 문제점을 개선하기 위한 DeiT[2], ConVit[3]와 같은 여러 모델들이 연구되고 있다.

따라서 본 논문에서도 많은 학습시간이 필요한 Transformer 비전 모델들의 문제점을 해결하기 위하여 Vision Transformer 분야의 모델 중 하나인 Co-Scale Conv-Attentional Image Transformer(CoaT)[4] 모델을 경량화한 Clustering CoaT 모델을 제안한다. Clustering CoaT 모델은 Adaptive Clustering module[5]을 적용하여 모델 학습 시 발생하는 계산 비용을 줄이고자 한다.

2. 관련연구

2.1 Co-Scale Conv-Attentional Image Transformers

CoaT 모델은 Transformer 기반의 이미지 분류 모델이다. 이 모델은 co-scale mechanism과 conv-attentional mechanism을 갖추고 있다. co-scale mechanism은 트랜스포머의 인코더 분기의 무결성을 유지하며 서로 다른 스케일로 학습한 표현을 효과적으로 전달할 수 있도록 한다. conv-attentional mechanism은 attentional 모듈에서는 효율적인 convolutional 유사 구현 방식을 사용하여 상대적 위치 임베딩 공식을 실현할 수 있도록 고안되었다. CoaT는 다양한 멀티스케일 및 상황별 모델링 기능을 통해 이미지 트랜스포머의 성능을 강화하였다. ImageNet에서 CoaT 모델은 유사한 크기의 CNN 및 이미지/비전 Transformer에 비해 우수한 분류 결과를 보여주었다[4].

2.2 End-to-End Object Detection with Adaptive Clustering Transformer 모델

End-to-End Object Detection with Adaptive Clustering Transformer 모델은 End-to-End Object Detection with Transformers (DETR)을 경량화한 모델이다. 이 모델은 DETR의 고해상도 입력에 대한 계산 비용을 줄이기 위해 Adaptive Clustering Transformer를 새롭게 제안하였다. Adaptive Clustering은 Locality Sensitive Hashing을 사용하여 Attention 레이어의 query 부분을 clustering 한다. Clustering 적용하게 되면 내부의 복잡성을 줄여줄 수 있다. 새롭게 적용한 Adaptive Clustering Transformer는 DETR과 비교하여 정확도와 계산 비용에서 좋은 결과를 보여주었다[5].

3. Clustering CoaT 모델 설계

Clustering CoaT 모델의 전체 흐름은 CoaT 모델의 흐름을 그대로 따른다. Clustering CoaT 모델은 CoaT-Lite 모델의 구성을 따라 다음과 같이 여러 Serial Block으로 구성되어 학습한다.

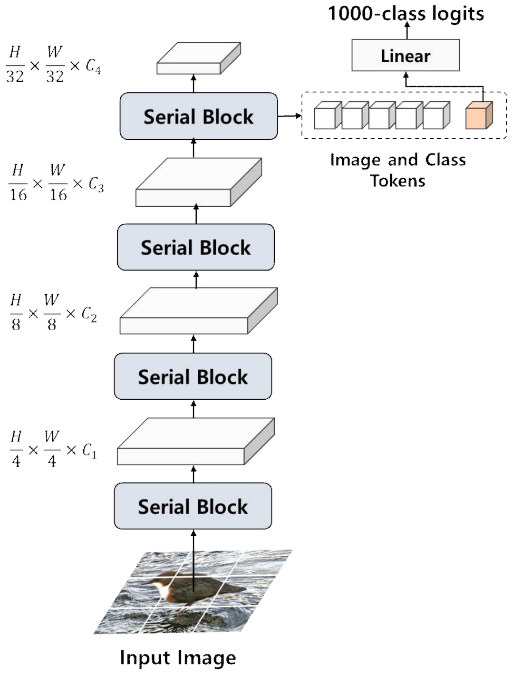

그림 1. CoaT-Lite model Architecture

Clustering CoaT 모델은 기존의 CoaT 모델의 conv-attentional module을 새롭게 수정하여 경량화를 하고자 한다. conv-attentional module에 Adaptive Clustering Attention module을 추가한다. 그림2는 Adaptive Clustering Attention module을 추가한 Clustering Conv-attentional module이다. conv-attentional module의 흐름은 다음과 같다. 입력된 feature map으로 depthwise conv를 계산한 후, 입력값에 더해준다. 그 후 나온 결과값을 value, query, keys로 나눠준다. 이때, 나눠진 query 값을 그대로 사용하는 것이 아닌 유사한 값들을 clustering 해준다. 그 후 clustering된 값들에서 대표값인 prototypes 값을 골라내어 key값과 계산하여 Attention map을 구해준다. Attention map은 value 값과 행렬곱하여 값을 구해주며, depthwise conv를 한 번 더 계산하여 value 값과 아다마르 곱을 해준 후 값을 더하여 output feature map을 구한다.

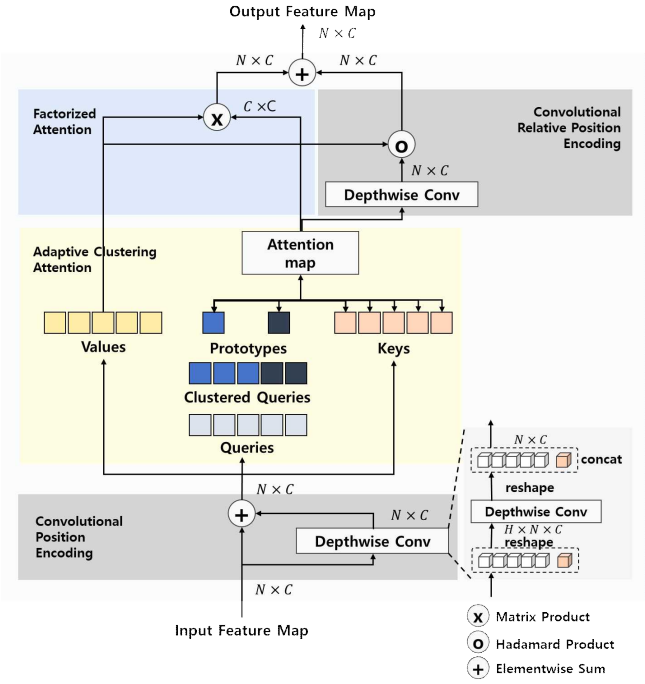

그림 2. Clustering Conv-attentional module Architecture

4. 결론

본 논문에서는 경량화를 위한 Clustering CoaT 모델을 제안하였다. Adaptive Clustering Attention module의 Clustering 기법을 활용하여 CoaT 모델의 내부 복잡도를 줄여 계산 비용을 낮추고자 하였다. 경량화된 Clustering CoaT 모델은 더 적은 학습시간과 학습데이터로 학습이 가능할 것을 기대한다. 이를 위해 향후 다른 모델과의 구체적인 비교 연구가 필요하다.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2021-2020-0-01489) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation). And this work was supported by the BK21 plus program through the National Research Foundation (NRF) funded by the Ministry of Education of Korea(5199990214660)

참고문헌

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

[2] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." International Conference on Machine Learning. PMLR, 2021.

[3] d'Ascoli, Stéphane, et al. "Convit: Improving vision transformers with soft convolutional inductive biases." International Conference on Machine Learning. PMLR, 2021.

[4] Xu, Weijian, et al. "Co-scale conv-attentional image transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[5] Zheng, Minghang, et al. "End-to-end object detection with adaptive clustering transformer." arXiv preprint arXiv:2011.09315, 2020.