DM-CLIP: Transformer에서 Mamba로의 Knowledge Distillation을 통한 효율적인 CLIP

조태완[1] · 노하림[2] · 유진희[3] · 양희재[4] · 최재용[1] †

가천대학교 AI·소프트웨어학부[1], 덕성여자대학교 컴퓨터공학전공[2],
고려대학교 통계학과[3], 성균관대학교 인공지능학과[4]

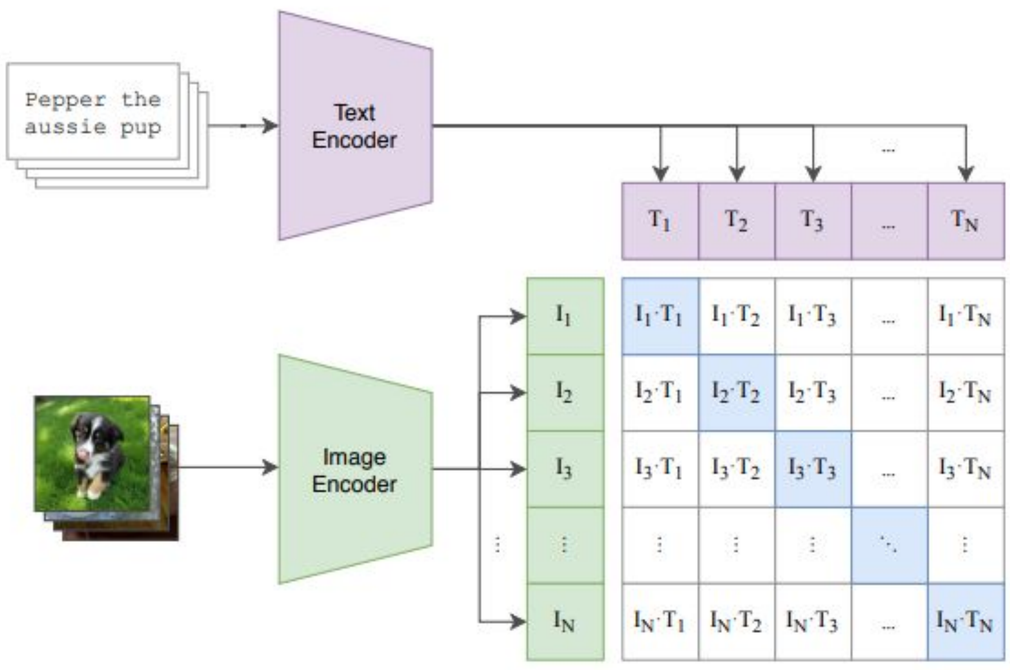
Abstract

- 본 연구는 이미지와 텍스트 간의 연관성을 학습하는 Contrastive Learning 기반 CLIP 모델의 높은 계산 복잡도와 큰 모델 크기가 리소스 제한 환경에서 활용을 어렵게 하는 문제를 해결하고자 함.
- 이를 위해 Transformer 기반 ViT 모델의 지식을 Mamba 기반 이미지 인코더로 Knowledge Distillation하여 성능을 개선하는 방식을 제안하고, 2가지 RQ를 설계하여 실험을 통해 이를 검증함.

Introduction

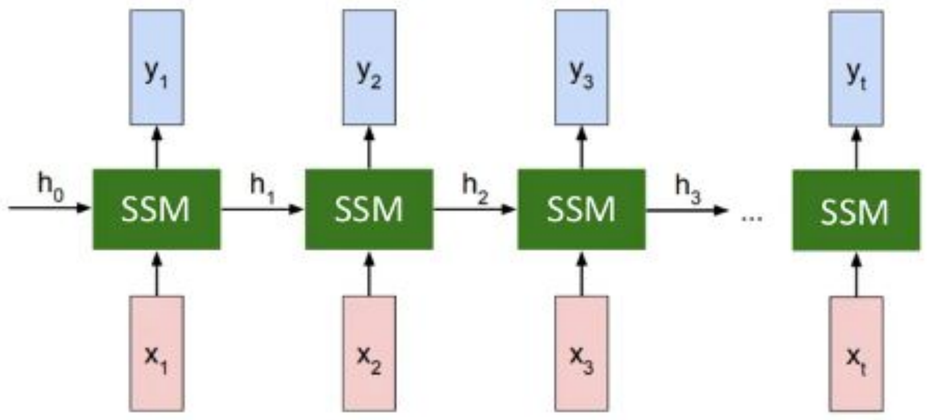
CLIP (Contrastive Language-Image Pre-training)

- 동일한 의미를 가지는 이미지-텍스트가 유사한 임베딩을 가지도록 학습시키는 방법.
- 대규모 데이터셋으로 학습시켜 강력한 zero-shot 성능을 가지고 있다는 것이 특징임.
- Self-Attention 메커니즘으로 인해 고해상도 이미지 처리 시 높은 연산량 요구.



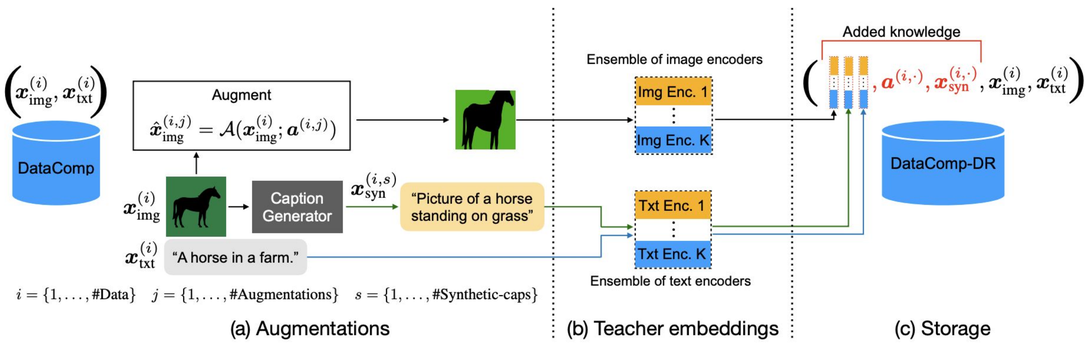
Mamba

- State Space Model (SSM)을 기반으로 한 시퀀스 모델링 기법으로, 선형적 시간 복잡도를 제공하여 기존 Transformer의 quadratic complexity 문제를 해결
- Self-attention 대신 효율적인 SSM 연산을 활용하여 긴 시퀀스 데이터를 효과적으로 처리할 수 있음(고해상도 이미지 처리가 용이함).
- 기존 연구에서 Hessian landscape의 복잡성으로 인해 global optimum 도달이 어려워 학습에 제약이 있을 수 있음을 주장함.



MobileCLIP

- 강화된 데이터셋인 DataCompDR과 Transformer 기반 Teacher 임베딩을 활용하여 효율적인 모델을 만든 연구
- 본 연구에서 DataCompDR을 활용하여 ViT 임베딩을 Mamba 기반 이미지 인코더로 Knowledge Distillation하여 기존의 문제를 해결할 수 있는지 확인하고 효율성을 검증함.



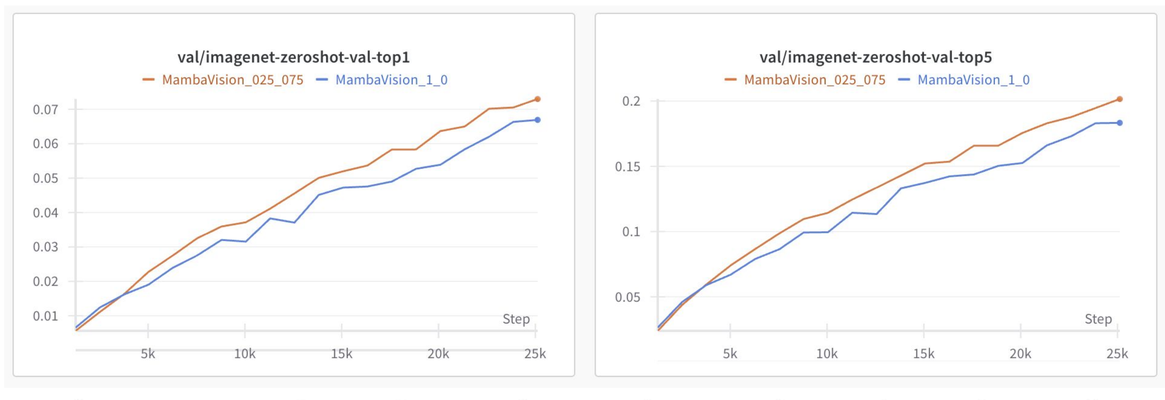
Results

Research Question

- RQ1: Mamba기반의 CLIP 인코더를 Contrastive Learning으로만 학습시키는 것 보다 Transformer 기반 모델의 Teacher Embedding을 Knowledge Distillation을 하는 것이 효과가 있는가?
- RQ2: Transformer기반 모델보다 Mamba기반 모델로 제작한 CLIP 인코더가 얼마나 효율적이며 어떤 데이터셋에서 강점을 보이는가?

RQ1 분석

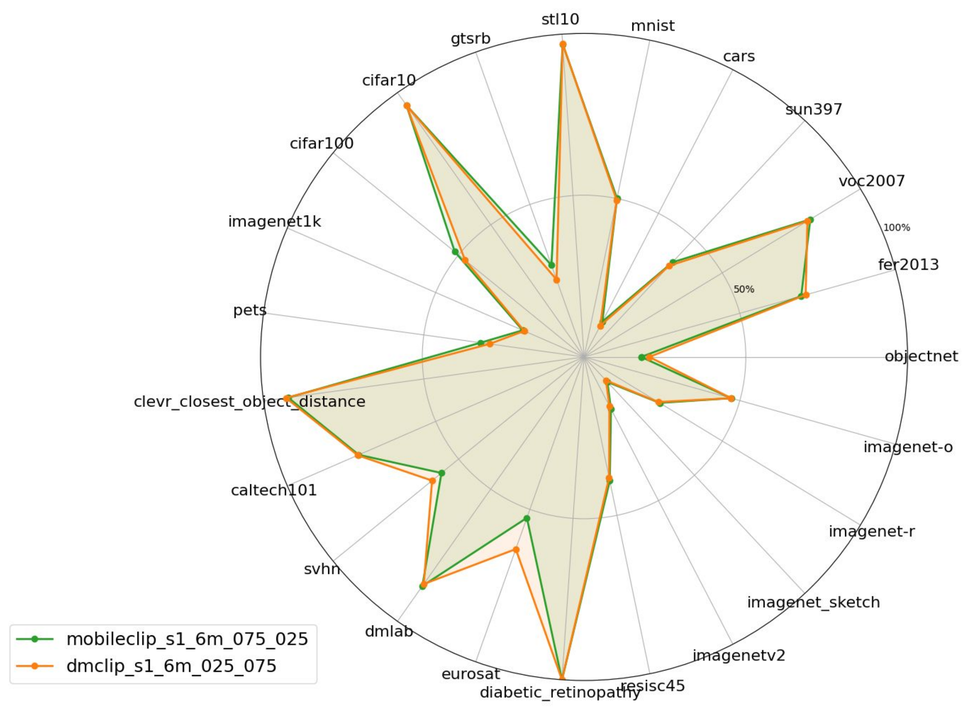
- Distillation 없이 Contrastive Loss 가중치를 1로 둔 모델(파란색)에 비해, Distillation Loss 가중치가 0.75이고 Contrastive Loss 가중치가 0.25인 모델(주황색)이 Acc1과 Acc5 모두 약 10% 향상. 다양한 평가지표에서도 좋은 성능을 보여줌.



Model	IN-val	IN-shift	Flickr30k T→I	Flickr30k I→T	COCO T→I	COCO I→T	Avg Perf. 23
dmclip_s1_6m_1_0	18.32	23.92	20.96	27.9	13.94	18.8	50.81
dmclip_s1_6m_025_075	**20.14**	**25.25**	22.72	26.0	**15.18**	20.44	**52.33**
dmclip_s1_6m_075_025	19.57	24.72	**23.36**	**30.4**	15.10	**20.76**	51.52
dmclip_s1_6m_0_1	20.00	24.75	23.34	27.9	14.59	19.50	50.98

RQ2 분석

- DM-CLIP 이미지 인코더의 지연 시간(latency)은 MobileCLIP 대비 49.58% 감소. 반면 성능 저하는 0.12%에 불과하며, 성능 손실 없이 큰 효율성 확보.
- 연속적 숫자 패턴 및 잡음이 많은 SVHN 데이터셋에서 정확도 약 6.6% 증가. 고해상도 위성 이미지인 EuroSAT에서 정확도 약 19.4% 증가.(acc5)



Conclusion

- 본 연구는 Mamba 기반 CLIP 인코더의 효율성과 성능을 두 가지 연구 질문(RQ1, RQ2)을 통해 분석함.
- 결론적으로, Mamba 기반 CLIP 인코더는 우수한 성능과 효율성을 동시에 제공하며, 특히 고해상도 데이터 처리 및 연속 패턴 학습이 필요한 응용에서 강력한 가능성을 보여줌.
- 더 큰 규모의 데이터셋과 다양한 배치 크기를 통해 Mamba 기반 인코더의 성능 한계를 검증할 예정