Ko-ViLT: Vision-Language Transformer 에 기반한 한국어 시각 질의응답

서민택 나승훈
전북대학교

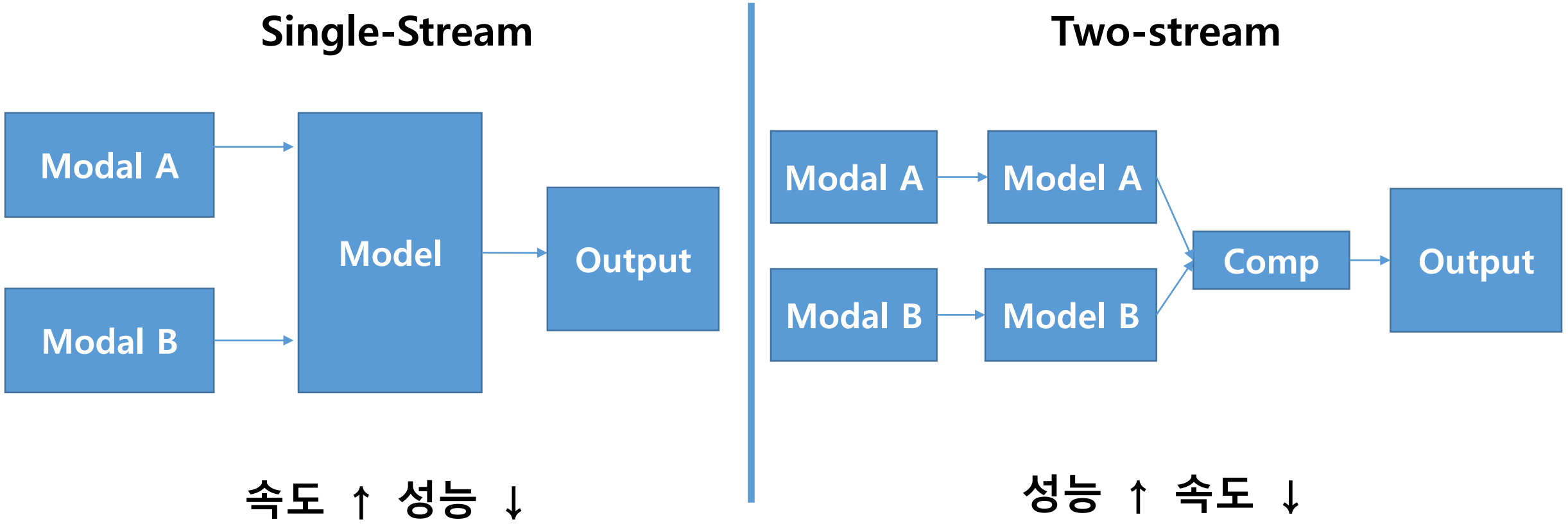
서대룡 김선훈 강인호
네이버

KCC22

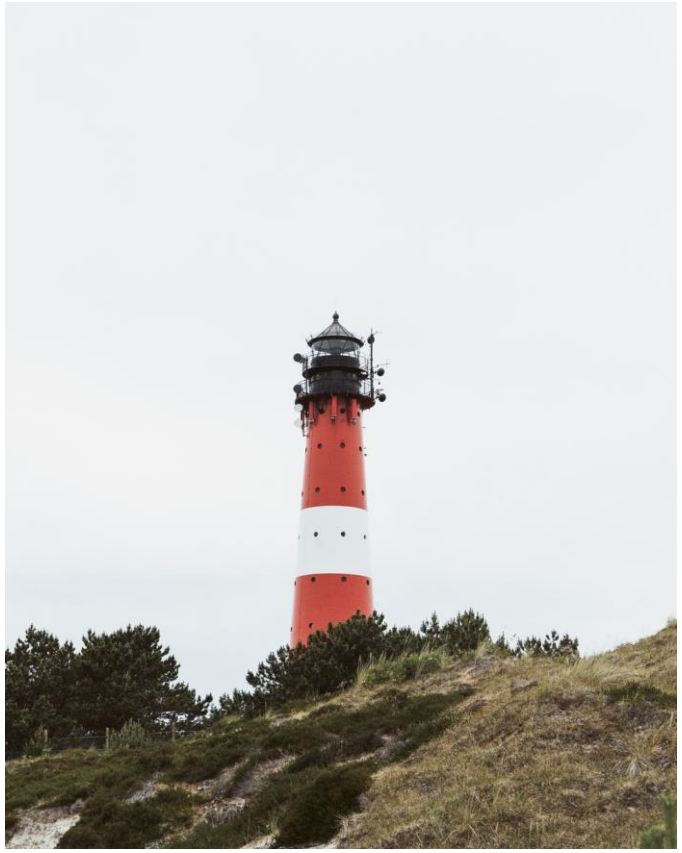
Introduction

- 자연어처리에서는 Transformer 구조를 통해서 높은 성능을 달성
- 최근 Vision에서도 Transformer 구조를 통해서 여러 Task에서 높은 성능을 달성함
- 각자의 단일 작업에서 여러 작업에 대해 높은 성능을 달성하여 최근 두 개 이상의 특성을 이해하는 Multimodal 작업에 대해 많은 연구 진행 됨.
- 본 논문에서는 Vision-Text Multimodal model을 한국어로 학습하고 모델의 성능에 대해 소개

Single-Stream vs Two-stream



MultiModal
Image embedding method



Object Detection
Region Feature

CNN
Backbone

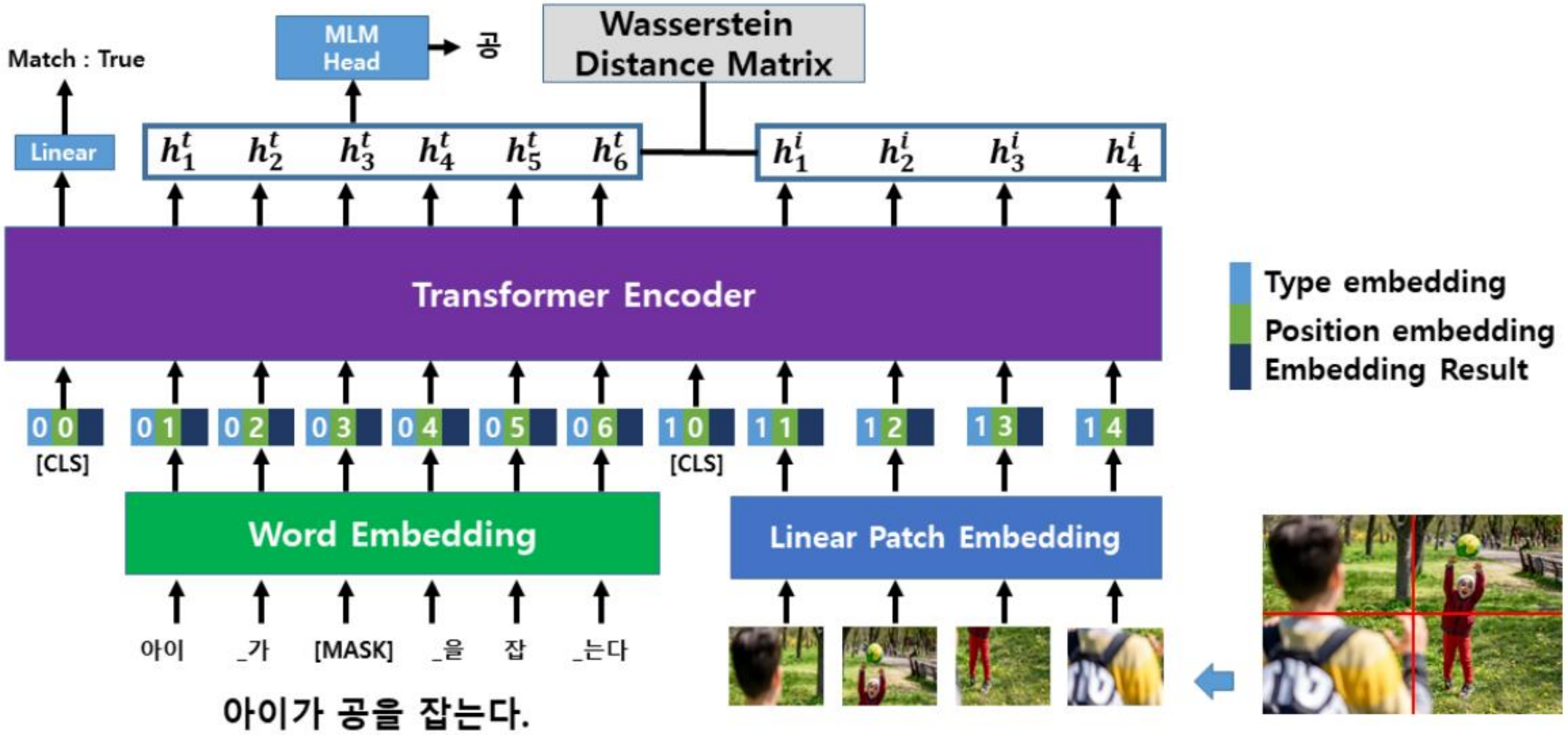
Linear Layer
Embedding

Image
Embedding
Tensor

많은 연산
낮은 속도

적은 연산
높은 속도

ViLT(Vision-Language Transformer)



Pretrained Loss

- [MASK] 혹은 임의 대치한 단어를 맞추는 Masked Langauge loss
- 이미지와 텍스트가 주어질 때 두 쌍이 연관된 쌍인지 판별하는 ITM(ImageTextMatching) loss
- ITM loss를 계산할 때 일치하는 쌍에 대해서 Text 은닉상태와 Image 은닉상태의 확률 분포 거리가 일치시에는 최소화되도록 불일치시에는 최대화 되도록 학습하는 최적 운송 문제 해결

한국어 시각 질의 응답

질의 : 사진의 소파는 무슨 색인가요?

- 이미지와 텍스트 모두 고려해야하는 Multimodal 작업

성능 비교

표 2: VQA 성능 비교

Test	Acc	1 Epoch Time
METER	85.24	24H
w/o pretrained ViLT	79.46	2H
w/ pretrained ViLT	78.61	2H

표 3: 사전학습 유무에 따른 수렴 속도 차이

Test	Epoch1	Epoch2	Epoch3	Epoch4	Epoch5
w/o pretrained	5.21	5.21	5.21	5.21	5.28
w/ pretrained	5.21	5.21	5.21	9.14	21.95

결론

- 각 모델을 사용하는 Two-Stream보다 성능이 소폭 낮지만, 연산시간에 대해서 10배의 이득을 보여줌
- 사전학습이 성능이 더 낮지만 사전학습에 600K개의 데이터를 사용하고 Downstream 학습에 1.5M개의 데이터를 사용하여 높은 성능이 보장되지 않지만 수렴 속도에 대해 장점을 가짐
- 연산시간이 매우 낮아져 학교 연구실 단위에서 Multimodal 연구에 도움이 될 것으로 기대됨.