[논문 리뷰] VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

diddu · 방금 전

논문

💡 VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text_논문 리뷰

📌 Introduction

VATT 논문은 라벨이 부여되지 않은 비효율적인 visual 데이터에 대한 자기지도 학습 방법을 제안한다. 이를 위해 다양한 modalities에서의 self-supervised multimodal pretraining 모델을 연구하며, 비디오, 오디오, 텍스트를 입력으로 받는 VATT를 제안하고 modality-agnostic한 모델을 테스트하여 모델의 범용성을 탐구하고, 학습복잡도를 줄이는 DropToken 방법을 소개한다.

📌 Approach

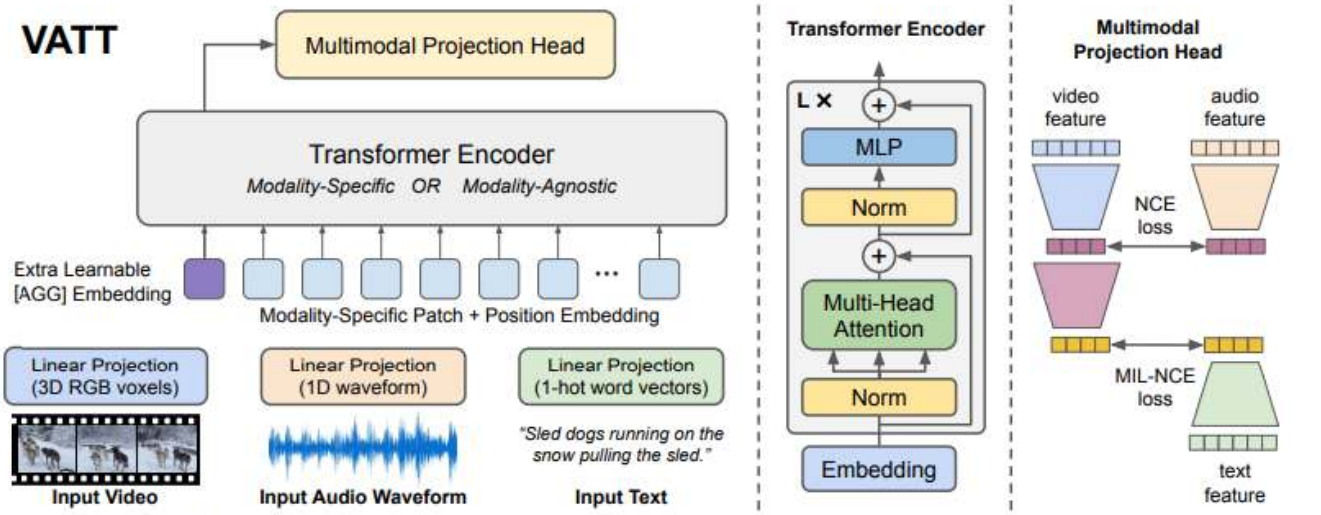


Figure 1. **Overview of the VATT architecture and the self-supervised, multimodal learning strategy**. VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities and employ the noise contrastive estimation to train the model.

`Tokenization & Positional Encoding`

각 모달리티의 데이터는 토큰화 및 위치 인코딩을 거쳐 트랜스포머 인코더의 입력으로 들어간다. 비디오 데이터는 패치로 분할되고, 오디오 데이터는 웨이브폼을 부분으로 나누며, 텍스트 데이터는 단어로 나누어진다. 이때, 위치 정보를 보존하기 위해 위치 인코딩이 사용된다.

`DropToken`

계산 복잡성을 줄이기 위해 비디오와 오디오 모달리티에서 입력 데이터의 일부를 임의로 삭제하는 DropToken 기법을 사용한다.

`Transformer Encoder`

각 모달리티의 입력에 위치 정보를 알려주는 positional encoding을 추가하여 트랜스포머 인코더를 구성한다. 이 모델은 모달리티별로 트랜스포머 인코더를 적용할 수 있으며, 두 가지 방법(Modality-Specific 및 Modality-Agnostic) 중 선택하여 사용한다.

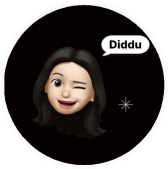
`Common Space` 매핑

모든 모달리티의 출력을 공통 공간으로 매핑한다. 비디오-오디오 쌍과 비디오-텍스트 쌍을 직접 비교할 수 있도록 의미적인 구조적인 공통 공간을 구축한다.

`Multimodal Contrastive Learning`
VATT 모델은 자기지도 목적함수를 사용하여 학습된다. 비디오-텍스트와 비디오-오디오 쌍을 비교하며, 이를 위해 Noise-Contrastive-Estimation(NCE)와 Multiple-Instance-Learning-NCE(MIL-NCE) 목적함수를 활용한다.

📌 Conclusion

- Transformer 아키텍처를 활용하여 self-supervised multimodal representation learning을 수행하는 프레임워크를 다루고 있다.
- 자기 지도 학습을 통해 다중 모달리티 데이터의 표현을 개선하고, 이를 통해 다양한 downstream 작업에서 우수한 성능을 달성함.
- DropToken 방법을 사용하여 연산량 문제를 완화하고, 효율적인 모델을 구축함.



Diddu

이전 포스트
[논문 리뷰] Tacotron 2 : NATURAL TTS SYNTHESIS BY CONDITIONING WAVE...

0개의 댓글

댓글을 작성하세요

댓글 작성


Powered by
Stellate