VATT: Video-Audio-Text Transformer

https://arxiv.org/abs/2104.11178

0. Introduction

- VATT는 영상(Video), 오디오(Audio), 텍스트(Text) 세 가지 모달리티를 동시에 처리할 수 있는 트랜스포머 기반 모델임
- 기존 연구들은 각 모달리티별로 CNN이나 RNN 기반 모델을 따로 사용함
- 이런 접근 방식은 모달리티 간 관계를 충분히 학습하지 못하고, 모달리티별 구조가 달라 확장성과 통합 학습에 한계가 있음
- VATT는 모달리티별 특화된 구조 없이 모든 입력을 동일한 트랜스포머로 처리함
- 모델 설계의 핵심은 컨볼루션 계층 없이 순수 트랜스포머만으로 멀티모달 표현을 학습하는 것
- 라벨 없는 대규모 데이터에서도 자가 지도 학습(self-supervised learning)을 통해 의미 있는 표현(feature)을 얻을 수 있음
- 목적 :
    - 세 가지 모달리티를 하나의 통합된 표현 공간(shared embedding space)에 학습
    - 모달리티 간 상호 이해(cross-modal understanding) 향상
    - 다양한 다운스트림 작업(video recognition, audio classification, text-video retrieval)에서 성능 개선

1. Overview

- VATT는 영상, 오디오, 텍스트를 동일한 트랜스포머 구조로 처리하는 멀티모달 모델임
- 각 모달리티 입력은 원시(raw) 형태 그대로 사용됨
    - 영상은 프레임 단위 픽셀
    - 오디오는 스펙트로그램으로 변환
    - 텍스트는 토큰화(tokenization) 후 임베딩

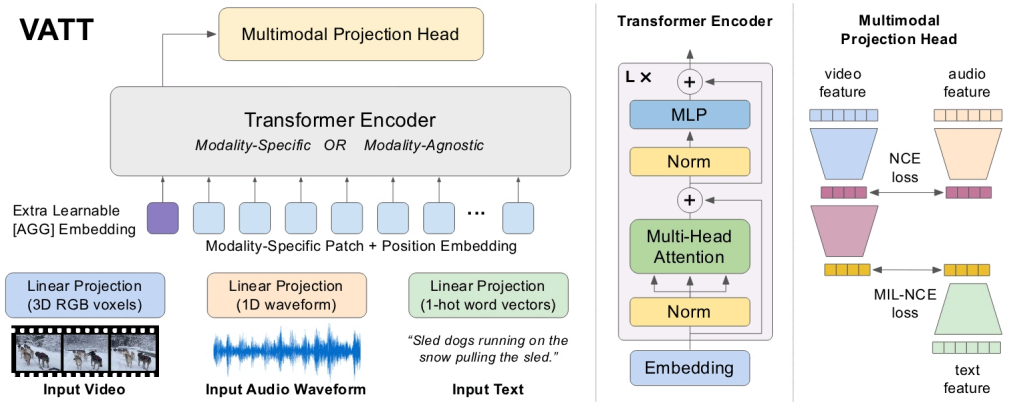
- 모델 구조는 모달리티별로 별도 인코더를 두지만 트랜스포머 구조와 하이퍼파라미터는 공유됨
- 모달리티 간 관계 학습 위해 대조 학습(contrastive learning) 적용
  - 같은 샘플의 다른 모달리티 표현은 가까이 다른 샘플은 멀리 배치하도록 학습
- 트랜스포머 어텐션 메커니즘을 활용해 시공간적(spatiotemporal) 관계를 캡처함
- 주요 장점 :
  - 모달리티별 CNN/RNN 불필요, 모델 구조 단순
  - 모달리티 간 상호작용 학습 가능
  - 라벨 없는 대규모 데이터 활용 가능
- 다운스트림 작업에 맞춰 파인튜닝 없이도 바로 적용 가능
  - 영상 분류(Video Classification)
  - 오디오 이벤트 분류(Audio Event Classification)
  - 텍스트-영상 검색(Text-Video Retrieval)
- 기존 멀티모달 학습 모델과 차별점 :
  - 단일 트랜스포머 구조로 모든 모달리티 처리
  - 컨볼루션 없이 순수 어텐션 기반
  - 라벨 없는 데이터에서 효과적인 자가 지도 학습 가능

2. Challenges

- 멀티모달 학습 자체가 어려움
  - 서로 다른 모달리티는 표현 방식, 시간적·공간적 특성이 달라 직접 결합 어려움
- 기존 CNN/RNN 기반 멀티모달 모델 한계
  - 모달리티별 특화 구조 필요
  - 모달리티 간 상호작용 학습이 제한적
  - 학습 및 추론 시 복잡한 구조, 높은 계산 비용 발생
- 라벨 없는 데이터 활용 어려움
  - 멀티모달 자가 지도 학습 설계 복잡

    - 서로 다른 모달리티 간 대조 학습 설계 필요
- 확장성과 일반화 문제
    - 기존 모델은 새로운 모달리티 추가 시 구조 변경 필요
    - 다양한 다운스트림 작업에 쉽게 적용하기 어려움
- 대규모 데이터 처리 문제
    - 영상, 오디오, 텍스트 모두 고차원 데이터이므로 메모리·연산 부담 큼
    - 시공간적 특성 유지하면서 효율적으로 트랜스포머 학습 설계 필요

3. Method



- 영상, 오디오, 텍스트 원시 데이터를 동일한 트랜스포머 구조로 처리함
- 각 모달리티별 입력 전처리
    - 영상 : 프레임 단위 픽셀 → 토큰화 후 패치 임베딩
    - 오디오 : 스펙트로그램 변환 → 패치 임베딩
    - 텍스트 : 토큰화 → 임베딩 벡터
- 모달리티별 별도 인코더 존재하지만 트랜스포머 구조와 하이퍼파라미터 공유
- 컨볼루션 계층 없이 순수 어텐션 메커니즘만 사용
- 모달리티 간 관계 학습 위해 대조 학습(contrastive learning) 적용
    - 같은 샘플 내 다른 모달리티 표현은 가까이
    - 다른 샘플 표현은 멀리 배치

- 학습 시 자가 지도 학습(self-supervised learning) 적용, 라벨 없는 데이터 활용 가능
- 시공간적 정보 유지 위해 트랜스포머에 위치 인코딩(position encoding) 적용
- 다운스트림 작업별로 추가적인 파인튜닝 없이도 사용 가능
- 모달리티 간 표현 통합 후 공유 임베딩 공간(shared embedding space) 생성

4. Experiments

- 영상 데이터 : Kinetics-400, Kinetics-600, Kinetics-700, Moments in Time 사용
- 오디오 데이터 : AudioSet 사용
- 이미지 데이터 : ImageNet 사용
- 텍스트-영상 검색 데이터 : MSR-VTT 사용
- 각 모달리티 입력은 원시 데이터 그대로 트랜스포머에 입력
- 동일 트랜스포머 구조로 학습
- 모달리티별 인코더와 하이퍼파라미터 공유
- 모달리티 간 관계 학습 위해 대조 학습(contrastive learning) 적용
- 학습 시 자가 지도 학습(self-supervised learning) 활용
- 라벨 없는 데이터에도 적용 가능
- 실험 환경 :
  - GPU 기반 분산 학습 사용
  - 배치 사이즈 및 학습률 최적화
- 다운스트림 작업별 성능 평가 :
  - 영상 분류(Video Classification)
  - 오디오 이벤트 분류(Audio Event Classification)
  - 텍스트-영상 검색(Text-Video Retrieval)

5. Results

Method	Kinetics-400 Top-1	Kinetics-400 Top-5	Kinetics-600 Top-1	Kinetics-600 Top-5	Moments in Time Top-1	Moments in Time Top-5	TFLOPs
I3D [13]	71.1	89.3	71.9	90.1	29.5	56.1	-
R(2+1)D [26]	72.0	90.0	-	-	-	-	17.5
bLVNet [27]	73.5	91.2	-	-	31.4	59.3	0.84
S3D-G [96]	74.7	93.4	-	-	-	-	-
Oct-I3D+NL [20]	75.7	-	76.0	-	-	-	0.84
D3D [83]	75.9	-	77.9	-	-	-	-
I3D+NL [93]	77.7	93.3	-	-	-	-	10.8
ip-CSN-152 [87]	77.8	92.8	-	-	-	-	3.3
AttentionNAS [92]	-	-	79.8	94.4	32.5	60.3	1.0
AssembleNet-101 [77]	-	-	-	-	34.3	62.7	-
MoViNet-A5 [47]	78.2	-	82.7	-	39.1	-	0.29
LGD-3D-101 [69]	79.4	94.4	81.5	95.6	-	-	-
SlowFast-R101-NL [30]	79.8	93.9	81.8	95.1	-	-	7.0
X3D-XL [29]	79.1	93.9	81.9	95.5	-	-	1.5
X3D-XXL [29]	80.4	94.6	-	-	-	-	5.8
TimeSFormer-L [9]	80.7	94.7	82.2	95.6	-	-	7.14
VATT-Base	79.6	94.9	80.5	95.5	38.7	67.5	9.09
VATT-Medium	81.1	**95.6**	82.4	96.1	39.5	**68.2**	15.02
VATT-Large	**82.1**	95.5	**83.6**	**96.6**	**41.1**	67.7	29.80
VATT-MA-Medium	79.9	94.9	80.8	95.5	37.8	65.9	15.02

Method	mAP	AUC	d-prime
DaiNet [21]	29.5	95.8	2.437
LeeNet11 [55]	26.6	95.3	2.371
LeeNet24 [55]	33.6	96.3	2.525
Res1dNet31 [49]	36.5	95.8	2.444
Res1dNet51 [49]	35.5	94.8	2.295
Wavegram-CNN [49]	38.9	96.8	2.612
VATT-Base	**39.4**	**97.1**	**2.895**
VATT-MA-Medium	39.3	97.0	2.884

Table 2: Finetuning results for AudioSet event classification.

Method	Pre-training data	Top-1	Top-5
iGPT-L [16]	ImageNet	72.6	-
ViT-Base [25]	JFT	**79.9**	-
VATT-Base	-	64.7	83.9
VATT-Base	HowTo100M	78.7	93.9

Table 3: Finetuning results for ImageNet classification.

Method	Batch	Epoch	YouCook2 R@10	YouCook2 MedR	MSR-VTT R@10	MSR-VTT MedR
MIL-NCE [59]	8192	27	**51.2**	**10**	**32.4**	**30**
MMV [1]	4096	8	45.4	13	31.1	38
VATT-MBS	2048	4	45.5	13	29.7	49
VATT-MA-Medium	2048	4	40.6	17	23.6	67

Table 4: Zero-shot text-to-video retrieval.

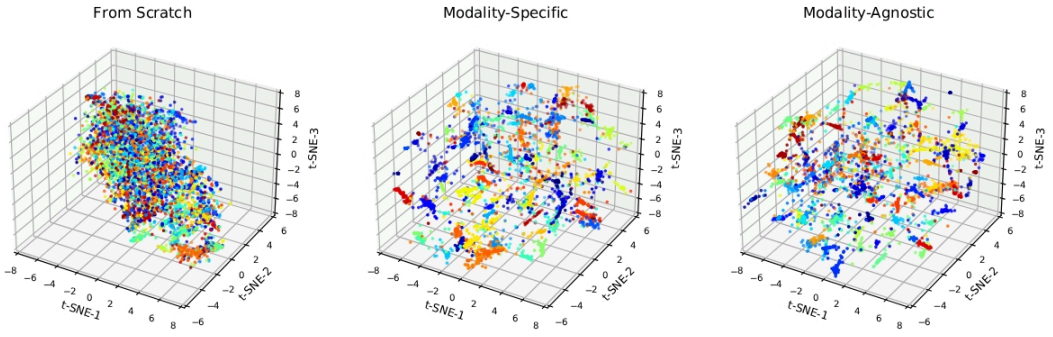
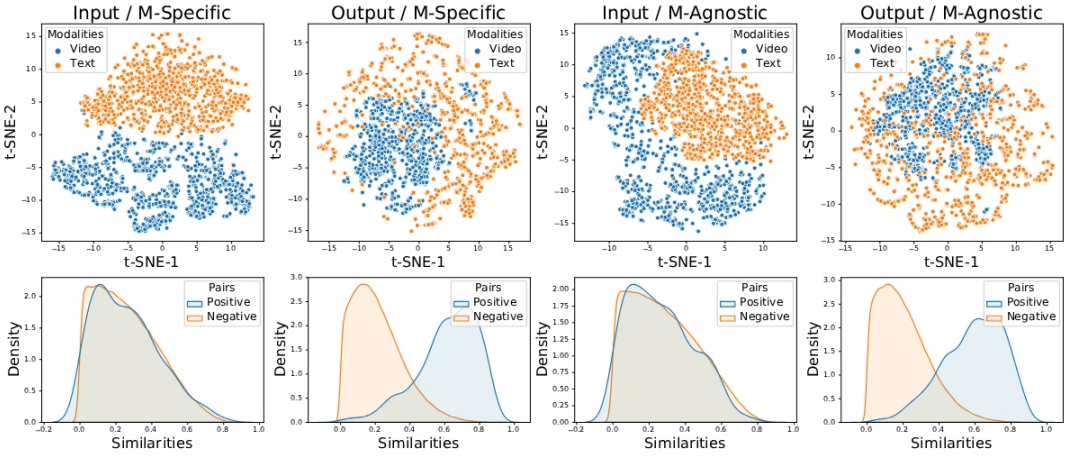
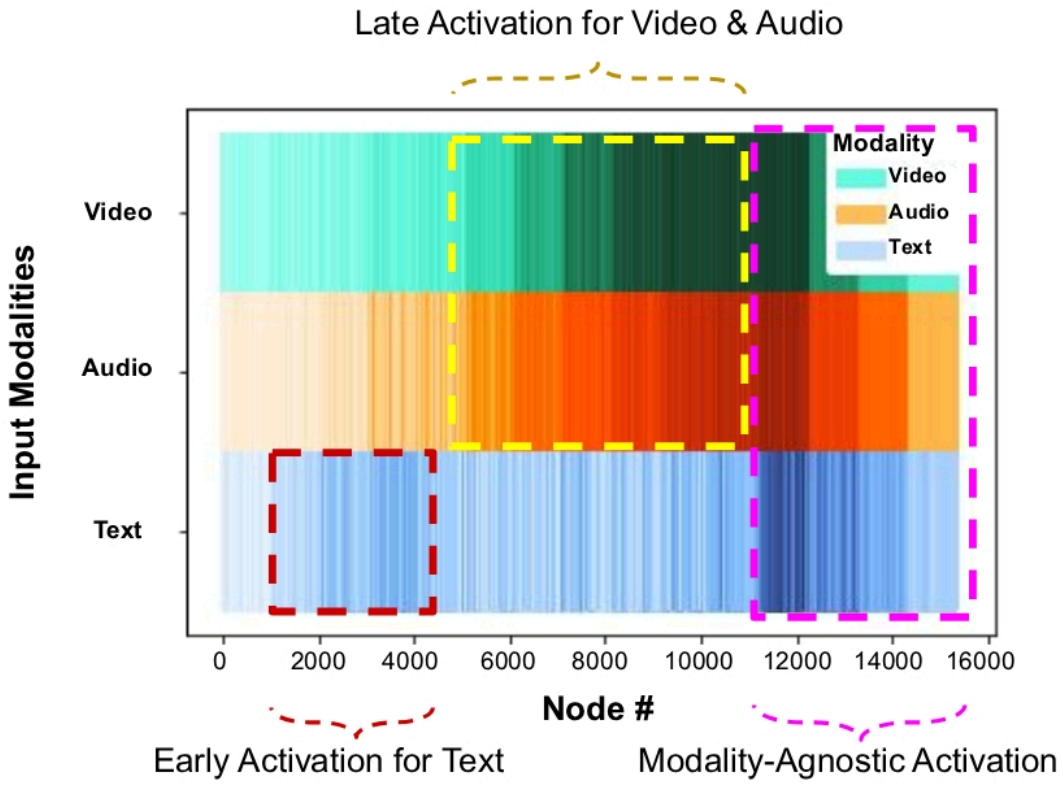


Figure 2: t-SNE visualization of the feature representations extracted by the vision Transformer in different training settings. For better visualization, we show 100 random classes from Kinetics-400.





	DropToken Drop Rate			
	75%	50%	25%	0%
Multimodal GFLOPs	188.1	375.4	574.2	784.8
HMDB51	62.5	64.8	65.6	66.4
UCF101	84.0	85.5	87.2	87.6
ESC50	78.9	84.1	84.6	84.9
YouCookII	17.9	20.7	24.2	23.1
MSR-VTT	14.1	14.6	15.1	15.2

Resolution/ FLOPs	DropToken Drop Rate			
	75%	50%	25%	0%
$32 \times 224 \times 224$	-	-	-	79.9
Inference (GFLOPs)	-	-	-	548.1
$64 \times 224 \times 224$	-	-	-	80.8
Inference (GFLOPs)	-	-	-	1222.1
$32 \times 320 \times 320$	79.3	80.2	80.7	81.1
Inference (GFLOPs)	279.8	572.5	898.9	1252.3

- 영상 인식 성능 :
  - Kinetics-400 Top-1 정확도 82.1%
  - Kinetics-600 Top-1 정확도 83.6%
  - Kinetics-700 Top-1 정확도 72.7%
  - Moments in Time Top-1 정확도 41.1%
- 이미지 분류 성능 : ImageNet Top-1 정확도 78.7%
- 오디오 이벤트 분류 성능 : AudioSet mAP 39.4%
- 텍스트-영상 검색 성능 : MSR-VTT에서 성능 향상 관찰
- VATT는 CNN 기반 모델 대비 영상·오디오 분야에서 우수한 성능 달성
- 단일 트랜스포머 구조로 모달리티 간 상호작용 효과적 학습
- 라벨 없는 데이터에서도 높은 표현 학습 능력 보여줌
- 다운스트림 작업별 추가 파인튜닝 없이도 적용 가능
- 멀티모달 통합 학습으로 모달리티 간 시너지 효과 확인

6. Insight

- 동일 트랜스포머 구조로 영상, 오디오, 텍스트를 통합 학습함
- 모달리티 간 관계를 효과적으로 캡처하여 멀티모달 표현 학습에 성공
- 라벨 없는 대규모 데이터에서도 자가 지도 학습으로 의미 있는 표현 획득 가능
- 컨볼루션 없이 순수 어텐션 기반 구조로 단순하면서도 확장성 뛰어남
- 학습된 표현은 다운스트림 작업에 바로 활용 가능, 파인튜닝 부담 감소
- 멀티모달 통합 학습이 기존 CNN/RNN 기반 접근보다 효율적임 확인
- 텍스트-영상 검색, 영상/오디오 분류 등 다양한 작업에서 성능 향상 관찰

- 모델 구조 단순화와 모달리티 통합으로 향후 새로운 모달리티 추가도 용이함
- 멀티모달 학습의 일반화 가능성과 실용성 입증