



2. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

0. Abstract

- **비디오-오디오-텍스트 트랜스포머 (VATT)**
 - 레이블이 없는 데이터를 활용하여 다중 모달 표현을 학습하는데 사용되는 컨볼루션 없는 트랜스포머 아키텍처
 - 비디오, 오디오, 텍스트를 입력으로 받아서 이들 간의 다중 모달 표현을 추출
- **VATT**를 처음부터 끝까지 다중 모달 대비 손실을 사용하여 학습하고, 비디오 액션 인식, 오디오 이벤트 분류, 이미지 분류, 텍스트에서 비디오 검색과 같은 작업에서 성능을 평가
- 세 가지 다른 모달(→ 비디오, 오디오, 텍스트) 간에 가중치를 공유하는 **모달-중립 트랜스포머**를 연구
- VATT는 기존의 **ConvNet** 기반 아키텍처보다 더 우수한 성능을 보임
 - 비디오 분야에서는 Kinetics-400, Kinetics-600, Kinetics-700 및 Moments in Time과 같은 데이터셋에서 뛰어난 성과를 보임
 - 이미지 분류 작업에서도 탁월한 일반화 능력을 나타냄
 - 오디오 이벤트 인식 작업에서도 새로운 기록을 세우고 있음

1. Introduction

- 해당 연구는 이전의 컴퓨터 비전 작업에서 성공을 거둔 합성곱 신경망(**CNNs**)과 자연어 처리(**NLP**)에서의 **트랜스포머 아키텍처**의 성과를 고려
 - **CNNs**: 시각 데이터에 대한 효과적인 귀납적 바이어스를 가짐
 - **NLP**: 트랜스포머 아키텍처가 강력한 성능을 발휘하고 있음

⇒ 대규모 지도 훈련은 레이블이 지정된 데이터가 많이 필요하고 시간과 비용이 많이 들기 때문에 제한적임

- 해당 연구에서는 비디오, 오디오 및 텍스트 모달리티를 포함한 원시 데이터를 입력으로 사용하는 트랜스포머 아키텍처인 **VATT**를 개발하고 이를 다양한 컴퓨터 비전 작업에서 평가함
 - 다중 모달 비디오로부터 사전 훈련되며, 레이블이 지정되지 않은 대규모 시각 데이터로 훈련됨
- 결과
 - **VATT**는 이미지 분류, 비디오 액션 인식, 오디오 이벤트 분류 및 텍스트-비디오 검색과 같은 다양한 작업에서 우수한 성과를 보임
 - 또한, 비디오 및 오디오 모달리티 사이에서 가중치를 공유하는 모달리티 중립 트랜스포머도 유망한 결과를 얻었음
 - 훈련 복잡성을 줄이는 효과적인 **DropToken** 기술도 소개

⇒ 트랜스포머 아키텍처가 다양한 데이터 유형에 대해 범용적으로 사용 가능한 다재다능한 모델임을 입증

⇒ 대규모 지도 훈련을 최소화하여 비지도학습 데이터(= unlabeled data) 활용의 중요성 강조

2. 관련 연구

2-1. 시각(비전) 분야에서의 transformer

- 트랜스포머는 원래 NLP 작업을 위해 설계됨
 - multi-head attention의 디자인은 단어 간의 장기 상관 관계를 모델링하는 데 효과적임
 - **Transformer**를 이미지 초해상도, 객체 탐지, 그리고 멀티모달 비디오 이해와 같은 비전 작업에 사용하려는 시도가 있었음
 - 이러한 방법들은 여전히 CNNs에 의해 추출된 특징에 의존
 - 최근에는 컨볼루션이 없는 비전 트랜스포머 집합을 제안하여 원시 이미지에서 직접 작동하고 CNNs와 경쟁력 있는 성능을 얻었음
 - 더 강력한 데이터 증강 및 지식 증류를 사용하여 이전의 훈련 데이터 효율성을 개선하였음

- 그 이후로 순수한 트랜스포머 디자인이 의미 분할, 포인트 클라우드 분류, 액션 인식과 같은 다양한 비전 작업에 적용됨

⇒ **VATT**는 비디오, 오디오 및 텍스트의 원시 멀티모달 입력에 대한 첫 번째 트랜스포머 모델임

2-2. 자기 지도 학습

자기 지도 학습

- 레이블이 지정되지 않은 데이터로부터 유용한 표현을 학습하는 방법 → 비전 작업에서 자주 사용됨
- 초기 자기 지도 학습 작업은 주로 이미지에 대한 사전 텍스트 작업을 수행하여 진행되었음
 - 자동 인코딩, 패치 위치 예측, 직소 퍼즐 해결, 이미지 회전 예측 등이 포함되었음
- 최근에는 대조적 학습이라는 접근 방식이 더 인기를 얻고 있음
 - 비디오 도메인에서는 주로 시간 정보를 활용하여 사전 텍스트 작업을 수행
 - 미래 프레임 예측, 움직임 및 외관 통계, 속도 및 인코딩 예측, 프레임 또는 비디오 클립 정렬이 포함됨
 - 최근에는 비디오에 대한 대조적 학습도 적용되고 있음
 - 시간 샘플링 전략과 일관된 시간적 증강을 사용하여 수행됨
 - 데이터 증강과 인스턴스 구별을 결합하여 이미지와 해당 증강된 뷰 간의 일관성을 유지

멀티모달 비디오

- 비디오는 다중 모달 데이터의 자연적인 요소임
 - 비디오가 오디오 스트림과 대응 관계를 가지는지 예측
 - 모드 간 클러스터링 및 진화하는 손실을 활용
- 최근에는 비디오, 오디오 및 텍스트로부터 학습하기 위해 대조적 손실을 사용하거나 좁은 시각에서 더 긴 시간적 컨텍스트를 포괄하는 넓은 시야를 예측하는 연구 등이 진행됨

⇒ **VATT**는 컨볼루션 없는 트랜스포머와 다중 모달 대조적 학습의 장점을 결합하는 첫 번째 작업

3. 접근

- 각 모달리티를 토큰화 레이어에 입력
 - 원시 입력은 임베딩 벡터로 투영된 다음 트랜스포머에 입력됨
- 주요 구성 요소
 1. backbone transformer
 - 모달에 따라 별도로 존재
 - 각 모달리티에 대한 특정 가중치를 포함하고 있음
 2. 단일 backbone transformer
 - 가중치 공유 → 어떤 모달리티에도 적용 가능
 - 어느 설정에서도 모달리티(= task) 별 표현을 추출
 - 추출된 표현은 대조적 손실을 사용하여 서로 비교할 수 있도록 공통 공간으로 매핑됨

3-1. 토큰화와 위치 임베딩

- VATT 아키텍처는 비전, 오디오, 텍스트 모달리티의 입력을 처리
 - 비전: 3채널 RGB 비디오 프레임으로 구성
 - 오디오: 공기 밀도의 진폭 정보(웨이브폼)를 포함
 - 텍스트: 단어 시퀀스로 이루어져 있음
- 각 모달리티에 대해 특정한 토큰화 레이어가 있음
 - 원시 입력을 임베딩 벡터로 변환한 후 트랜스포머에 입력됨
- 아키텍쳐의 주요 설정
 1. 각 모달리티에 대한 별도의 backbone 트랜스포머
 - 백본 트랜스포머가 분리되어 각 모달리티에 대한 특정 가중치를 가짐
 2. 가중치를 공유하는 백본 트랜스포머
 - 여러 모달리티에 동일한 백본 트랜스포머가 적용되며, 각 모달리티에 대한 토큰화 레이어와 선형 프로젝션 레이어가 별도로 존재
 - 모두 backbone transformer는 모달리티 별 고유 표현을 추출
→ 대조 손실을 통해 서로 비교하기 위해 공통 공간에 매핑됨

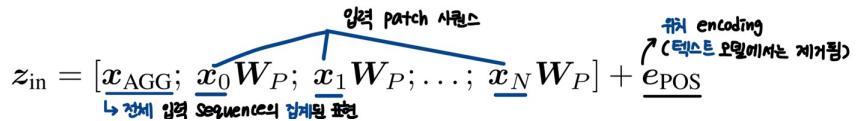
3-1-1. DropToken

- 훈련 중 계산 복잡성을 줄이기 위한 간단하면서도 효과적인 전략

- 비디오 또는 오디오 모달리티의 토큰 시퀀스를 얻으면 토큰 중 일부를 무작위로 샘플링하고, 샘플된 시퀀스를 전체 토큰 세트가 아닌 트랜스포머에 공급
→ 트랜스포머의 계산 복잡도가 입력 시퀀스의 토큰 수인 N에 대해 $O(N^2)$ 인 곳에서 효과적으로 계산 비용을 줄일 수 있음
 - 입력 길이를 줄이는 어떤 노력이든 연산 횟수(FLOPs)를 이차적으로 감소시킴
 - 모델 훈련 시간에 즉각적인 영향을 미침
 - 제한된 하드웨어에서 대형 모델을 호스팅하는 것을 가능하게 함
- ⇒ 원시 비디오와 오디오 입력과 같이 중복성이 높은 경우에 특히 효과적

3-2. Transformer의 구조

- NLP에서 널리 사용되는 Transformer 아키텍처 구조를 대부분 차용해 다른 표준 Transformer 구현에 가중치를 쉽게 전송할 수 있도록 함
 - 입력 토큰의 시퀀스



- 입력 패치 시퀀스 x_n
- 전체 입력 sequence의 집계된 표현인 x_{AGG}
→ 분류 및 공통 공간 매핑에 사용
- Multi-Head-Attention(MHA) 모듈: 표준 자기 어텐션을 사용
- MLP 레이어의 활성화 함수: GeLU
- MHA 및 MLP 모듈 앞에 Layer Normalization 을 사용
- 텍스트 모델에서는 위치 인코딩을 제거하고 첫 번째 MHA 모듈 레이어의 각 어텐션 점수에 대한 학습 가능한 상대적 편향을 추가
⇒ 텍스트 모델의 가중치를 최신 텍스트 모델 T5로 직접 전송 가능

3-3. 공통 공간 투영

- VATT(Video, Audio, Text Transformer)
 - 비디오, 오디오 및 텍스트 데이터를 다룸
 - 데이터 간의 의미론적 관련성을 학습하기 위해 공통 공간 투영과 대조 학습을 사용

- 네트워크는 비디오-오디오 쌍과 비디오-텍스트 쌍을 비교하여 공통 공간에 매핑하는 계층적인 접근 방식을 채택
 - 각 모달리티의 다양한 의미론적 세분화 수준을 고려하고, 이러한 차이를 규정하기 위해 **다중 수준 투영**을 사용

projection head

$$z_{v,va} = g_{v \rightarrow va}(z_{\text{out}}^{\text{video}}), \quad z_{a,va} = g_{a \rightarrow va}(z_{\text{out}}^{\text{audio}}) \quad \text{● : 선형 투영}$$

$$z_{t,vt} = g_{t \rightarrow vt}(z_{\text{out}}^{\text{text}}), \quad z_{v,vt} = g_{v \rightarrow vt}(z_{v,va}) \quad \text{● : ReLU를 활성화한 투영}$$

↑ The projection heads to map each modality into the common space.

- 비디오와 오디오를 공통 공간으로 매핑하는 투영 헤드
($g_{v \rightarrow va}, g_{a \rightarrow va}$)
- 텍스트와 비디오를 다시 공통 공간으로 매핑하는 투영 헤드
($g_{t \rightarrow vt}, g_{v \rightarrow vt}$)
- 선형 투영 및 ReLU 활성화 함수를 사용하여 공통 공간 투영을 정의
 - 각 선형 레이어 뒤에는 배치 정규화를 적용하여 훈련을 안정화시킵니다.



서로 다른 모달리티 간의 의미론적 차이를 고려하여 공통 공간으로의 투영을 조절함으로써 모델을 향상시키자!

3-4. 멀티모달 대조적 학습

- VATT는 비디오-오디오 쌍을 정렬하기 위해 Noise Contrastive Estimation (NCE)을 사용하고, 비디오-텍스트 쌍을 정렬하기 위해 Multiple Instance Learning NCE (MIL-NCE)을 사용
 - ⇒ 비디오-오디오-텍스트 스트림의 다른 시간적 위치에서 생성
- 모델은 양성 쌍을 생성하기 위해 비디오에서 동일한 위치에서 비디오와 해당 오디오를 샘플링하고, 음성 쌍을 생성하기 위해 비디오에서 비일치 위치에서 무작위로 샘플링함
- 공통 공간은 섹션 3에서 정의되며, 목표 함수는 양성 쌍과 음성 쌍을 구별하기 위해 온도 매개변수를 사용하여 정의됨

$$\text{NCE}(z_{v,va}, z_{a,va}) = -\log \left(\frac{\exp(z_{v,va}^\top z_{a,va}/\tau)}{\exp(z_{v,va}^\top z_{a,va}/\tau) + \sum_{z' \in \mathcal{N}} \exp(z'^\top_{v,va} z'_{a,va}/\tau)} \right), \quad (4)$$

$$\text{MIL-NCE}(z_{v,vt}, \{z_{t,vt}\}) = -\log \left(\frac{\sum_{z_{t,vt} \in \mathcal{P}} \exp(z_{v,vt}^\top z_{t,vt}/\tau)}{\sum_{z_{t,vt} \in \mathcal{P}} \exp(z_{v,vt}^\top z_{t,vt}/\tau) + \sum_{z' \in \mathcal{N}} \exp(z'^\top_{v,vt} z'_{t,vt}/\tau)} \right), \quad (5)$$

시전방에서 비디오 clip를
 가장 가까운 유사한 clip
 (양성 vs 음성 구분)
 (양성 vs 음성 구분)

두 원본 간의 균형 조정
 ↑ 비디오 대비
 ↑ 모든 비행기 영상

- 전체 모델의 학습 목표: 양성과 음성 쌍의 손실을 조절하는 하이퍼파라미터인 λ 를 사용하여 두 가지 손실 간의 균형을 유지하면서 배치 내의 샘플들의 평균 손실을 최소화하는 것

$$\mathcal{L} = \text{NCE}(z_{v,va}, z_{a,va}) + \lambda \text{MIL-NCE}(z_{v,vt}, \{z_{t,vt}\}),$$

두 원본 간의 균형 조정

4. 실험

4-1. 실험 설계

사전 학습

- AudioSet 및 HowTo100M 데이터셋을 결합하여 사전 훈련됨
 - 비디오-오디오-텍스트 삼중 샘플로 구성
 - 비디오) 32프레임을 10fps로 샘플링
 - 오디오) 48kHz에서 샘플링
 - 비디오와 오디오는 정규화되며, 텍스트는 one-hot encoding을 사용하여 처리됨
- 모델 훈련
 - DropToken 및 Adam 옵티마이저가 사용됨
 - 코사인 스케줄링된 학습률과 warmup 스텝이 적용됨
- 공통 공간 투영에는 특정 매개변수와 가중치가 설정됨
 - 다양한 네트워크 크기의 변형이 실험에 사용
- 시간: TPU 사용 → 상대적으로 짧음

downstream

- 사전 훈련된 VATT 모델을 총 10개 데이터셋을 사용하여 4가지 주요 하향식 작업에서 평가

- 비디오 액션 인식) UCF101, HMDB51, Kinetics-400, Kinetics-600, Moments in Time을 사용하여 평가
- 오디오 이벤트 분류) ESC50 및 AudioSet을 사용하여 평가
- 비디오-텍스트 공통 공간 표현) YouCook2 및 MSR-VTT에서의 zero-shot 텍스트-비디오 검색을 사용하여 평가
- 비전 백본의 전송성) ImageNet 분류에서 파인 투닝하여 평가
- HMDB51, UCF101 및 ESC50은 네트워크 크기와 비교하여 매우 작은 데이터셋
 - 사전 훈련된 백본 위에 선형 분류기를 훈련하는 용도로만 사용
- 선형 분류 정확도 및 제로샷 비디오 검색 메트릭을 보고

4-2. 실험 결과

4-2-1. 비디오 동작 인식을 위한 fine-tuning

- VATT 모델을 Kinetics-400, Kinetics-600 및 Moments in Time과 같은 대규모 동영상 액션 인식 데이터셋에서 파인 투닝하여 성능을 평가

METHOD	Kinetics-400		Kinetics-600		Moments in Time		TFLOPS
	TOP-1	TOP-5	TOP-1	TOP-5	TOP-1	TOP-5	
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
최신 모델 → TimeSFormer-L [9]	80.7	94.7	82.2	95.6	-	-	7.14
모델과의 비교	VATT-Base	79.6	94.9	80.5	95.5	38.7	67.5
	VATT-Medium	81.1	95.6	82.4	96.1	39.5	68.2
	VATT-Large	82.1	95.5	83.6	96.6	41.1	67.7
모델과의 우판	VATT-MA-Medium	79.9	94.9	80.8	95.5	37.8	65.9
							15.02

Table 1: Video action recognition accuracy on Kinetics-400, Kinetics-600, and Moments in Time.

- 해당 실험에서 네 가지 다른 사전 훈련 설정을 사용 → 최신 비디오 모델들과 비교
- 결과적으로 세 데이터셋 모두에서 TimeSFormer와 같은 최신 모델을 능가하는 높은 정확도를 달성
 - VATT의 사전 훈련은 인간이 정의한 레이블에 의존하지 않음

- 멀티모달 비디오 데이터에 대한 자체 지도 학습을 통해 사전 훈련된 최초의 비전 트랜스포머 백본으로 알려져 있음
- 또한, Kinetics-700 데이터셋에서도 VATT의 성능은 이전 연구를 뛰어넘어 상위 1% 정확도 72.7%를 달성했습니다.
- 멀티모달 자체 지도 학습 사전 훈련이 얼마나 도움이 되는지를 확인하기 위해 사전 훈련 없이 모델을 훈련시키고 낮은 정확도를 관찰함
- 마지막으로, VATT-MA-Medium이 비디오 액션 인식을 위해 파인 튜닝될 때 비디오, 오디오 및 텍스트 모달리티에서 공유되는 모달리티-무관 백본이 모달리티별 VATT-Base와 동등한 성능을 보이는 것으로 나타남

⇒ 단일 트랜스포머 backbone을 통합하여 세 가지 데이터 모달리티를 처리할 수 있는 잠재력을 보여줌

4-2-2. 오디오 사건 분류를 위한 fine-tuning

기준	METHOD	mAP AUC d-prime		
		곡선 평균 정밀도 아래 영역	AUC 기반	AUC 기반
기존 CNN 모델	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
성능 상승	VATT-Base 모달리티 병합 fine-tuning	39.4	97.1	2.895
	VATT-MA-Medium	39.3	97.0	2.884

Table 2: Finetuning results for AudioSet event classification.

- 오디오 이벤트 분류 작업에서 AudioSet 데이터셋을 사용하여 파인 튜닝
- 해당 실험에서는 두 가지 다른 사전 훈련 설정의 최종 체크포인트를 사용
 1. 모달리티별(BBS) transformer
 2. 모달리티-무관(Medium) transformer
- 결과
 - 일반적으로 사용되는 평가 메트릭인 평균 평균 정밀도 (mAP), 곡선 아래 영역 (AUC), 및 d-prime(→ AUC 기반)을 활용
 - VATT의 transformer는 모든 메트릭에서 기존의 CNN 기반 모델을 능가함

- 모달리티-무관 백본(VATT-MA-Medium)을 파인 튜닝하는 것이 모달리티-별 백본 (VATT-Base)을 파인 튜닝하는 것과 동등한 결과를 보임

4-2-3. 이미지 분류를 위한 fine-tuning

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.

- 멀티모달 비디오 도메인에서 사전 훈련된 모델을 사용하여 이미지 분류 작업에 지식을 전이할 수 있는 능력을 보여줌
- VATT-BBS의 비전 트랜스포머를 ImageNet에서 수정 없이 파인 튜닝하는 실험을 수행
 - 사전 훈련을 통해 처음부터 훈련하는 것과 비교하여 정확도가 크게 향상된 것을 확인 가능
 - 또한, 비지도 사전 훈련은 비디오 도메인에서 이루어지지만, 대규모 이미지 데이터를 사용한 지도 학습과 경쟁력 있는 결과를 달성한다는 것을 관찰 가능

4-2-4. zero-shot 텍스트 → 비디오 검색

METHOD	baseline	BATCH	EPOCH	YouCook2		MSR-VTT	
				R@10	MedR	R@10	MedR
MIL-NCE [59]		8192	27	51.2	10	32.4	30
MMV [1]	기본	4096	8	45.4	13	31.1	38
VATT-MBS	정도	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.

- 비디오-텍스트 쌍을 VATT-MBS에 공급하고 S_{vt} 공간에서 표현을 추출한 후, YouCook2 및 MSR-VTT의 각 비디오-텍스트 쌍 간 유사성을 계산
 - 텍스트 쿼리에 대해 유사성이 높은 비디오를 순위로 매겨서 상위 10개 비디오 중 올바른 비디오의 재현율을 측정하며, 올바른 비디오의 순위 중앙값도 측정
- 실험 결과**
 - 두 가지 베이스라인(MIL-NCE, MMV)과 비교
 - 배치 크기와 에포크 수가 검색 결과에 영향을 미치는 것을 확인
 - VATT**가 MMV와 유사한 결과를 제공하는 것을 확인

- 또한, 텍스트 transcript의 노이지한 특성
 - 고급 언어 모델이 과소평가되는 경향이 있다는 점을 언급
 - 미래 연구에서 더 높은 품질의 텍스트 소스 탐구의 필요성 제기

4-2-5. 특징 시각화

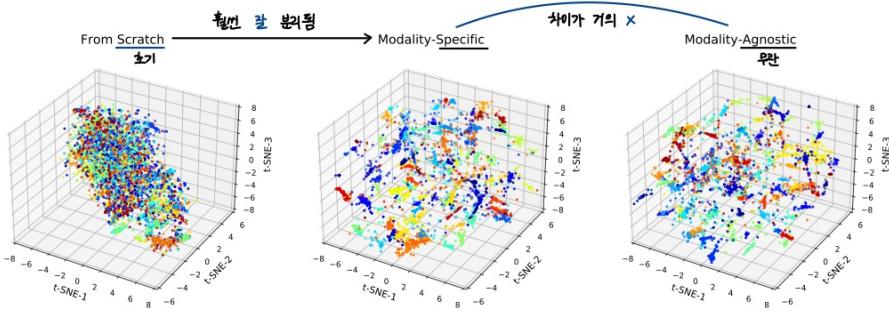


Figure 2: t-SNE visualization of the feature representations extracted by the vision Transformer in different training settings. For better visualization, we show 100 random classes from Kinetics-400.

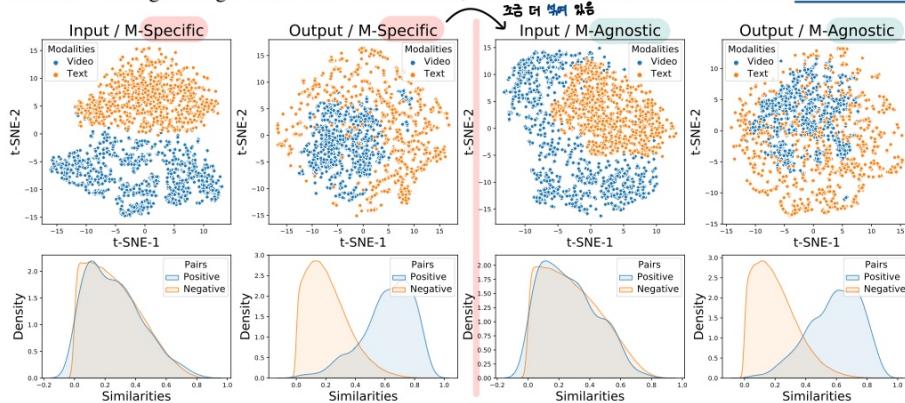


Figure 3: t-SNE visualization and distribution of pair-wise similarities of the input space vs. output space for modality-specific and modality-agnostic backbones when different modalities are fed.

- Kinetics-400**에서 fine tuning된 modality-specific 및 modality-agnostic VATT 모델의 출력 특성 표현을 t-SNE를 사용하여 시각화
 - 결과적으로 fine-tuning 된 VATT가 처음부터 훈련된 모델보다 훨씬 더 효과적으로 특성을 분리한다는 것을 확인
 - 또한, 모달리티-무관(modality-agnostic) 특성과 모달리티-별(modality-specific) 특성 간에 명확한 차이가 없다는 점을 관찰
- VATT backbone을 fine-tuning하지 않은 상태에서도 실험
 - YouCook2 데이터셋에서 선택한 비디오 클립에서 훈련된 VATT 모델의 두 지점에서 표현을 분석하고, modality-specific VATT와 modality-agnostic VATT를 비교
 - 모달리티-무관 설정에서 표현이 약간 더 혼합되어 있음을 관찰

⇒ 모달리티-무관 백본이 다른 모달리티를 동일한 개념을 나타내는 다른 기호로 인식한다는 것을 시사

- 또한, VATT가 양성 비디오-텍스트 쌍과 무작위로 샘플링한 쌍을 얼마나 잘 구별하는지 확인하기 위해 유사성 계산 & 시각화
⇒ VATT가 다양한 모달리티에 대한 의미론적 공통 공간을 학습하는 데 효과적임을 확인

4-2-6. 모델 활성화

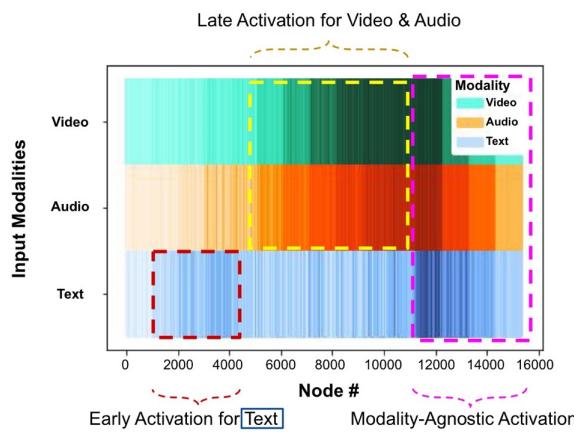


Figure 4: The average node activation across the Modality-Agnostic Medium VATT while feeding a multimodal video-audio-text triplet to the model.

- modality-agnostic VATT 모델에 전체 다중 모달 입력이 주어졌을 때의 평균 활성화를 측정
 - `HowTo100M`의 테스트 데이터셋에서 100,000개의 짧은 비디오 클립을 샘플링
→ 해당하는 오디오&텍스트와 함께 모델에 입력으로 공급
 - 각 모달리티에 대해 모델의 출력에서 MLP 모듈 이전의 각 노드의 평균 활성화를 계산
 - 결과
 - 초기 레이어의 노드: 텍스트 입력과 관련하여 활성화
 - 중간 및 이후 레이어의 노드: 비디오 및 오디오 모달리티와 관련하여 활성화
 - 네트워크의 마지막 레이어에 있는 노드: 거의 동일하게 모든 모달리티와 함께 활성화
- ⇒ 모델이 모든 모달리티에 대한 의미론적 인식을 달성하면서 나중 레이어에서 서로 다른 모달리티에 대한 노드를 할당할 수 있음을 시사

4-2-7. DropToken의 효과

	Drop Token 삭제 비율			
	Drop Token	Drop Rate	75%	50%
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

Table 5: Top-1 accuracy of linear classification and R@10 of video retrieval vs. drop rate vs. inference GFLOPs in the VATT-MBS.

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

Table 6: Top-1 accuracy of video action recognition on Kinetics400 using high-resolution inputs coupled with DropToken vs. low-resolution inputs.

- 고해상도 데이터에서의 중복을 줄이기 위한 DropToken 방법을 제안
 - 다운스트림 응용 프로그램 및 사전 훈련 계산에 미치는 영향을 연구하기 위해 비디오 및 오디오 입력에서 토큰을 무작위로 삭제하는 실험을 수행
 - 실험 결과, 50%의 토큰 삭제 비율이 정확도와 계산 비용 사이에서 좋은 균형을 제공하므로 대규모 사전 훈련에 이 비율을 선택하는 것이 적절하다는 것을 확인할 수 있음
 - 또한, 고해상도 입력과 DropToken을 결합한 경우가 훈련 정확도와 비용 면에서 저해상도 입력과 비교하여 유리하다는 결과를 제시
- ⇒ 저해상도 입력 대신 DropToken과 함께 고해상도 입력을 사용하는 것이 권장됨

5. 결론 & 논의

- 해당 논문에서는 Transformer를 기반으로 한 자체 지도 학습 다중 모달 표현 학습 프레임워크를 제안 ⇒ **VATT**
- 연구 결과**
 - Transformer가 비디오, 오디오, 텍스트와 같은 의미론적 표현을 효과적으로 학습하는 데 효과적임
 - 하나의 모델이 여러 모달리티를 공유하는 경우에도 다중 모달 자체 지도 사전 훈련이 대규모 라벨 데이터에 대한 의존성을 줄일 수 있음을 시사
 - DropToken이 비디오와 오디오 모달리티와 함께 사전 훈련 복잡성을 크게 줄일 수 있으며 모델의 일반화에 미치는 영향이 미미하다는 결과를 제시
- 해당 연구는 비디오 액션 인식 및 오디오 이벤트 분류에서 새로운 기록을 보이고 이미지 분류와 비디오 검색에서 경쟁력 있는 성능을 달성

⇒ 그러나 아직도 일부 제한 사항이 존재하며, 향후 연구에서 이러한 **제한 사항**을 개선할 필요가 있음

1. 모든 비디오에는 유기적인 오디오 또는 음성이 없지만, VATT의 접근 방식은 의미 있는 다중 모달 대응에 의존
2. 현재 텍스트 모달리티는 소음이 많고 때로는 희소한 음성 transcript로 구성되어 있음
3. 여전히 계산이 요구되는 면에서 우리의 방법은 요구가 있지만, 인간 라벨이 필요하지 않게 관리 ← 뭔소리야,,