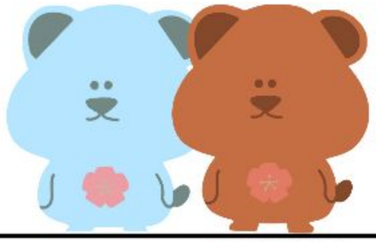
VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Euron 3주차 스터디
발표자 : 장윤서

EWHA EURON

목차



EWHA EURON

Introduction

EWHA EURON

#01 Introduction

컴퓨터 비전 분야에 attention 도입 시도

convolution과 attention 개념을 합한
하이브리드 모델의 제안

Attention

NLP → **CV**

Transformer

BERT

Vision Transformer

컴퓨터 비전
(이미지 처리)분야에
트랜스포머 도입

VATT

BERT, VIT 모델을 기반으로
영상 처리 분야까지
트랜스포머의 활용성을 확장

EWHA EURON

#01 Introduction

컴퓨터 비전 분야에 attention 도입이 어려웠던 이유

: 자체의 지도 능력(organic supervision)이 있는 자연어와 다르게 이미지 데이터는 지도 능력이 없기 때문

CNN과 Transformer 학습의 차이점

Natural language → Transformer

자연어는 순차적으로 단어, 구, 문장이 문맥 내에 놓여져 있고 이러한 순차적 배열을 통해 각각의 요소에 의미와 구문을 부여

Transformer 모델은 텍스트를 raw signal(레이블링 되지 않은 데이터)로 받아서 자기지도 학습을 진행할 수 있음

Image → CNN

이미지 픽셀값은 문맥이 없음, 객체의 모양이나 위치가 조금만 달라도 같은 객체인지 알 수 없다

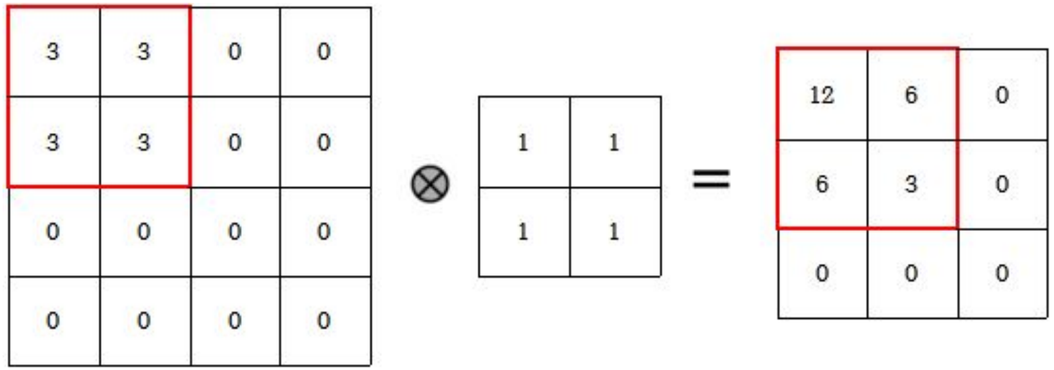
CNN, RNN은 이를 보완하기 위해 강한 inductive bias를 가짐

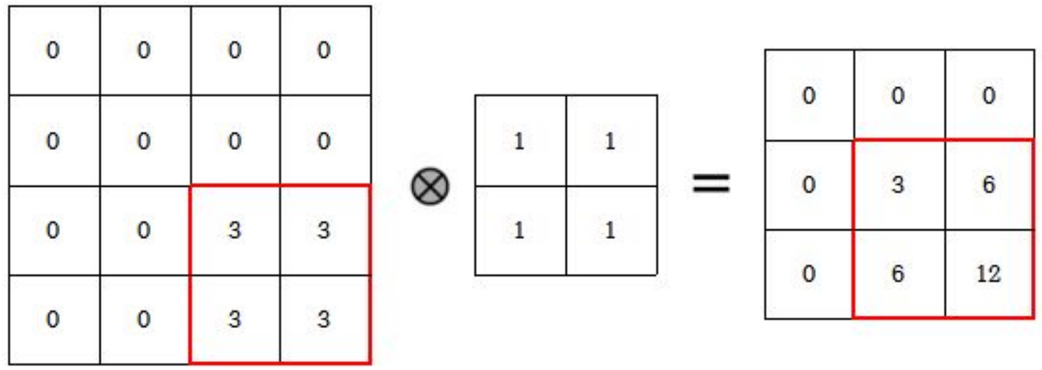
즉 보지 못한 데이터에 대해서도 귀납적 추론이 가능하도록 모델이 알고리즘 내에서 미리 가정하고 학습을 진행하는 것

#01 Introduction

컴퓨터 비전 분야에 attention 도입이 어려웠던 이유

: 자체의 지도 능력(organic supervision)이 있는 자연어와 다르게 이미지 데이터는 지도 능력이 없기 때문





CNN의 inductive bias의 간단한 설명

맥스풀링 레이어는 여러 픽셀 중 최댓값을 가진 픽셀 하나를 출력하기 때문에 객체 위치나 모양이 약간씩 다른 경우에도 동일한 객체로 인식할 수 있음

컨볼루션 연산에서 한 객체에 대해서는 같은 필터를 사용함으로써 같은 객체에 대한 파라미터를 공유, 강아지가 상단에 있던 하단에 있던 같은 강아지로 인식 할 수 있다

#01 Introduction

대규모 지도학습 기반 Transformer의 문제

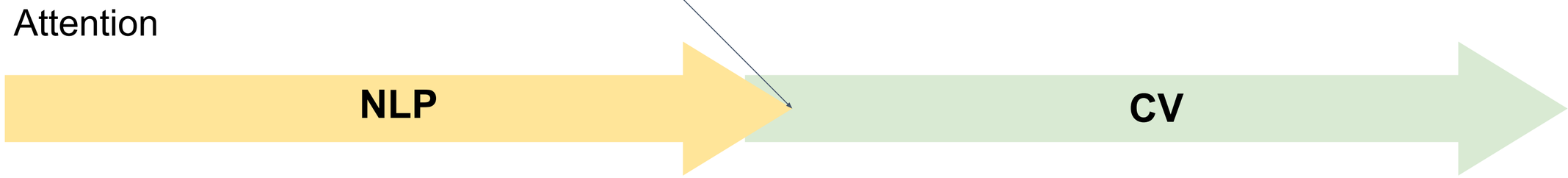
- CNN은 기본적으로 지도학습 모델

1. “big visual data”의 큰 부분을 배제
   (레이블이 없고 구조화되지 않은 데이터)
   이로 인해 모델이 어느 한쪽으로 편향되는 문제

2. 파라미터가 너무 많음, 소요 시간이 매우 크다

컴퓨터 비전 분야에 attention 도입 시도

convolution과 attention 개념을 합한
하이브리드 모델의 제안



#01 Introduction

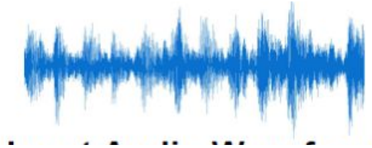
Video, Audio, Text Transformers(VATT) 모델 제안

멀티모달 영상에서의 자연어의 문맥과 비슷한 organic supervision의 특성 발견

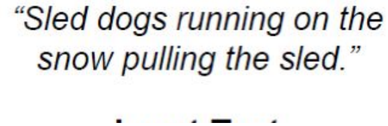
modality : 데이터 입력의 형식
멀티모달이란 두가지 이상의 센서로 수집된 데이터가 동시에 존재한다는 의미

영상에 포함된 3개 모달리티 :

영상



음성



,, 자막(텍스트)

영상, 음성, 텍스트 raw 데이터의 입력으로 사람의 라벨링 없이 Transformer 모델 학습을 진행할 수 있다는 가능성 제시

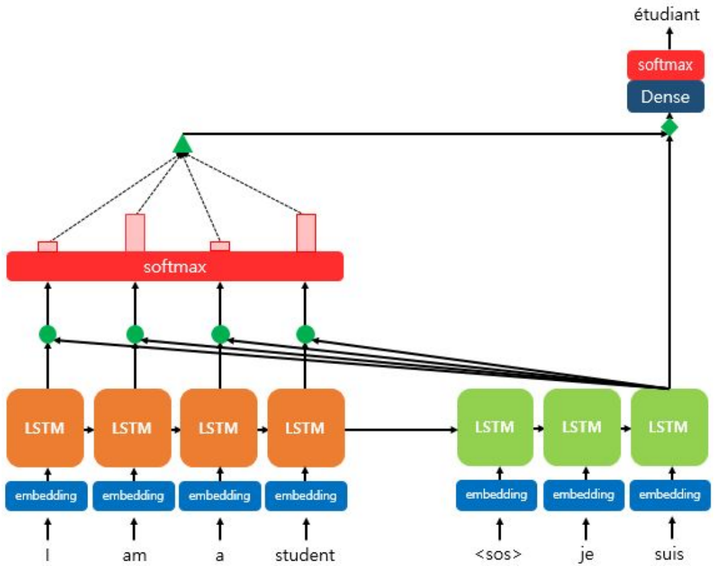
-> 아이디어가 재밌고 Transformer 활용 방안의 확장이기에 읽어보면 좋을 것 같아 해당 논문을 선택하게 됨

Related Work

#02 Related Work

Attention 모듈

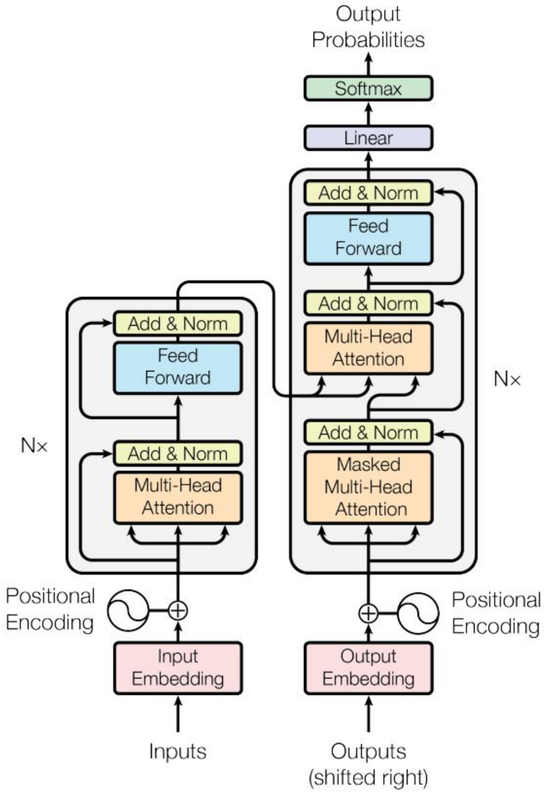
인풋의 각 토큰에 대해 문장의 토큰들이 얼마나 연관도를 가지는지를 weight로 계산헤 출력함으로써 문맥에 따라 집중할 단어를 결정할 수 있음



Transformer

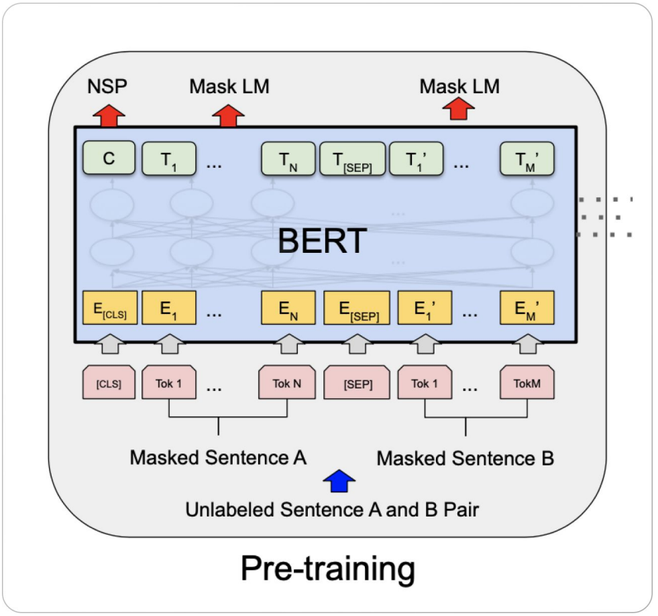
attention 연산을 여러번 진행함으로써 기존 RNN 모델보다 더 효율적으로 기계 번역 task 수행

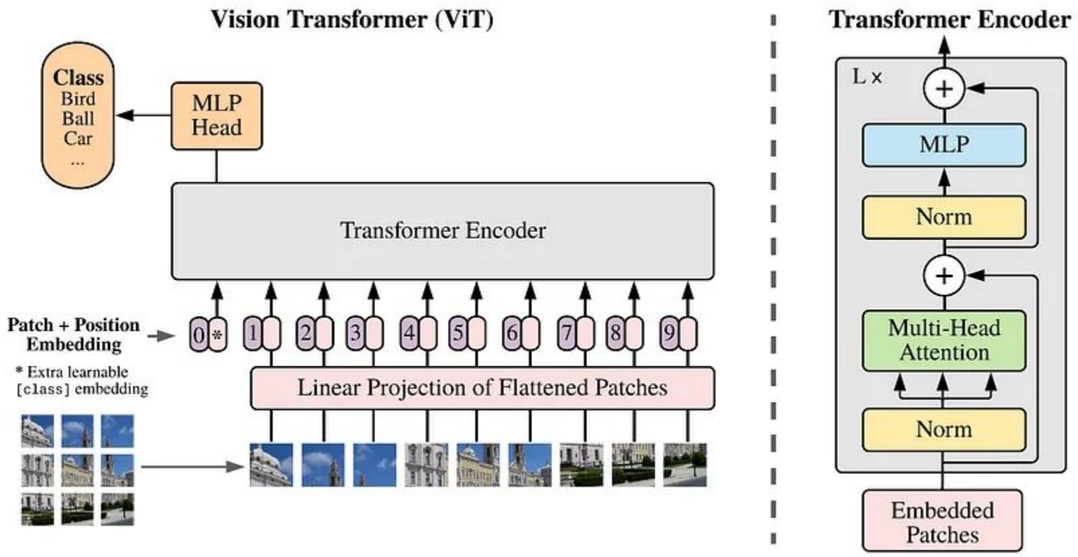
Attention 기법을 자연어 처리 분야에 제시함으로써 자연어 처리의 패러다임 변화를 가져옴



#02 Related Work

BERT, VIT(Vision Transformer) vs VATT





VATT의 모델 구조는 BERT, VIT의 모델과 유사하게 설계

modality별로 독립된 tokenization layer와 linear projection을 가진다는 차이점 외에는 거의 비슷

-> Transformer가 여러 종류의 데이터에 대해서도 범용성을 가지는 구조임을 밝힘

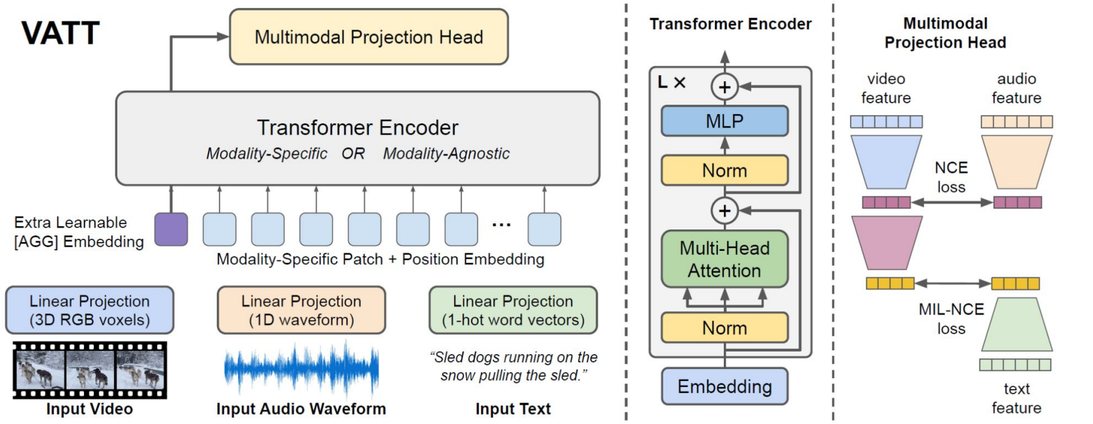
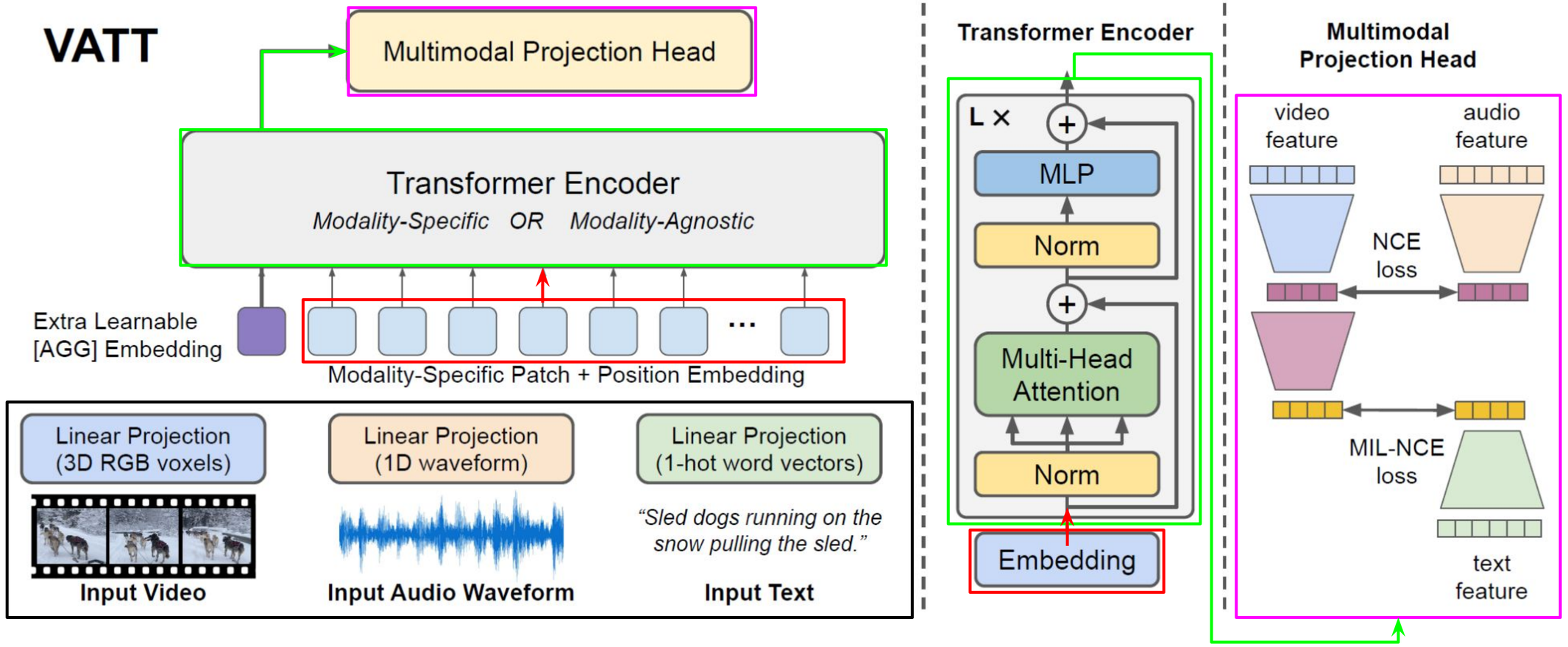


Figure 1. **Overview of the VATT architecture and the self-supervised, multimodal learning strategy**. VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities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

Model Architecture

#03 Architecture Overview

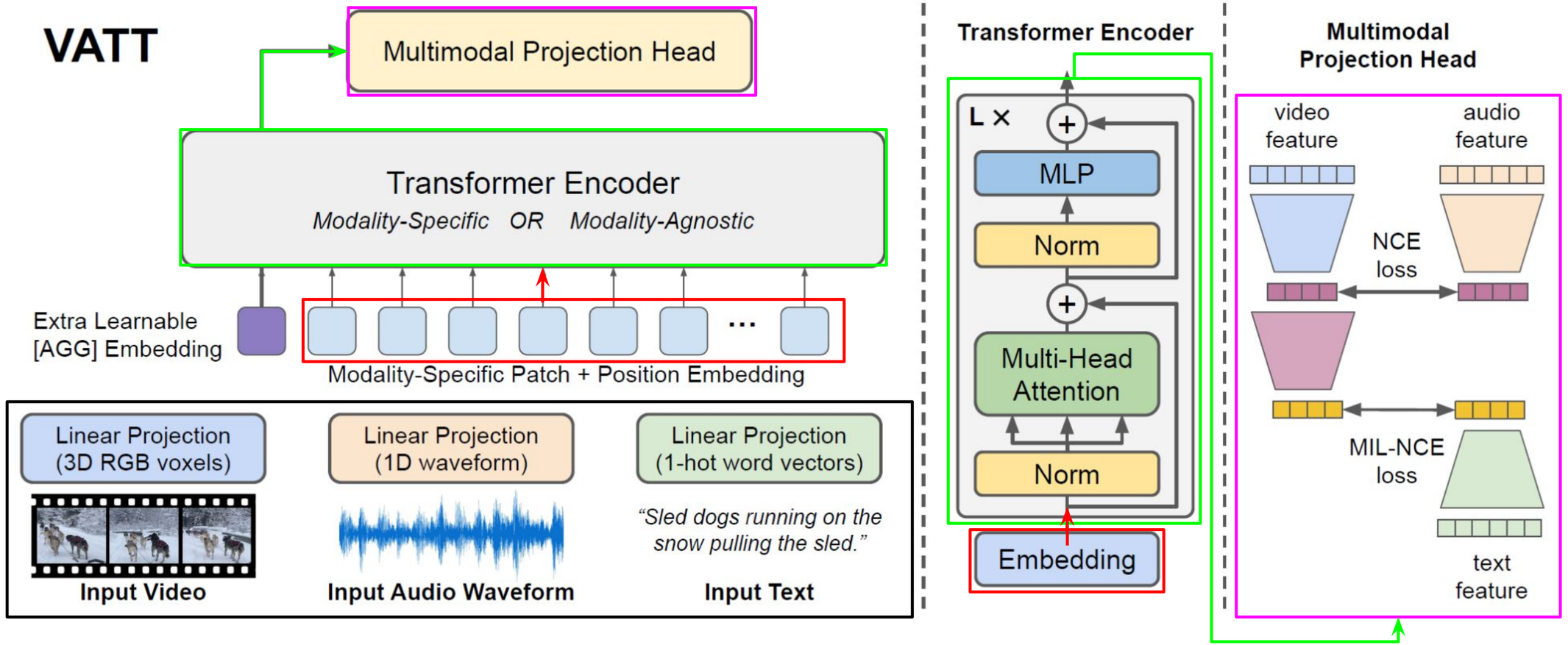
같은 색은 동일한 내용을 의미



1. Backbone Transformer가 분리되어 있으며 각 modality별로 고유의 weight를 갖는다
2. Transformer는 모든 modality에서 weight를 공유하는 단 하나만이 존재한다.

#03 Architecture Overview

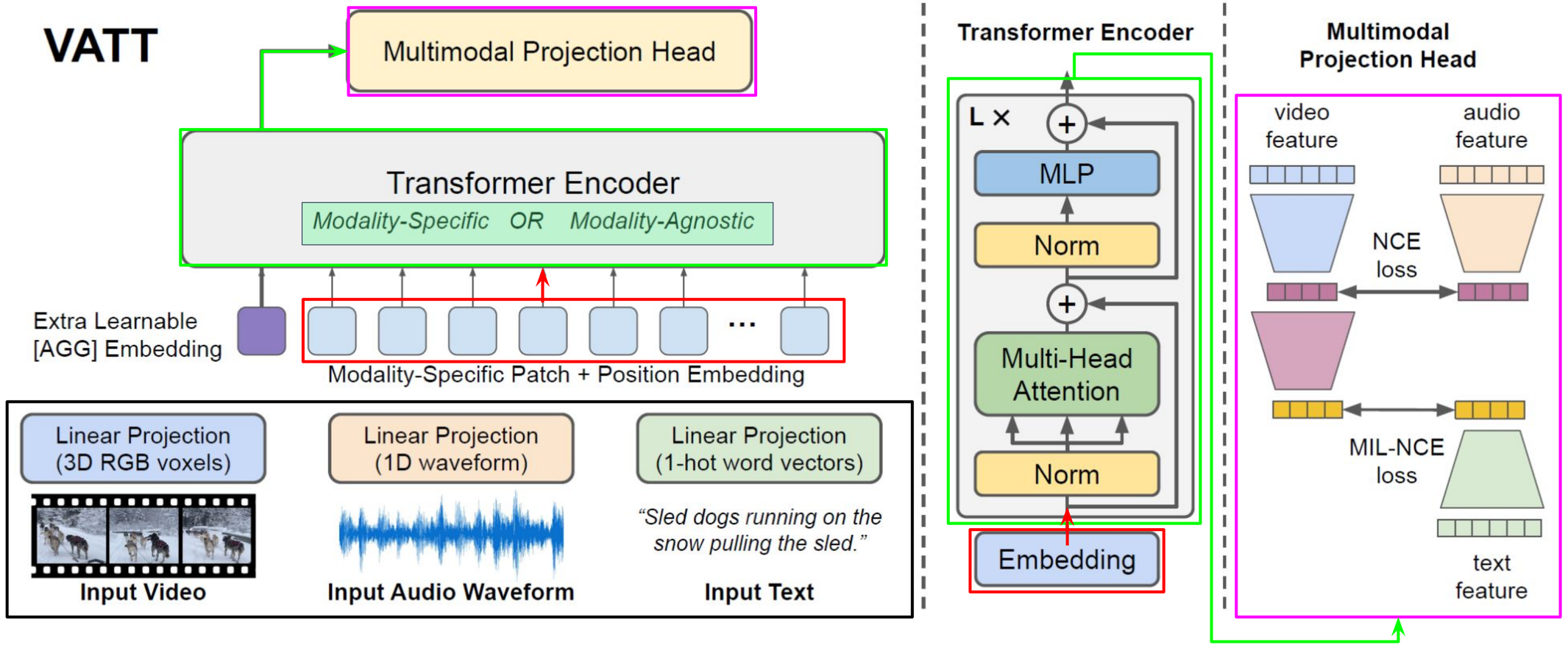
같은 색은 동일한 내용을 의미



RGB frame의 Video, Audio, one-hot word 벡터 형태인 Text 데이터가 각각 토큰화, Position Embedding을 거쳐 시퀀스를 가진 feature 벡터로 변환되고 트랜스포머 인코더로 입력됨

#03 Architecture Overview

같은 색은 동일한 내용을 의미



1. Modality-Specific

각 모달리티가 **독립적으로 트랜스포머 인코더를 거침**

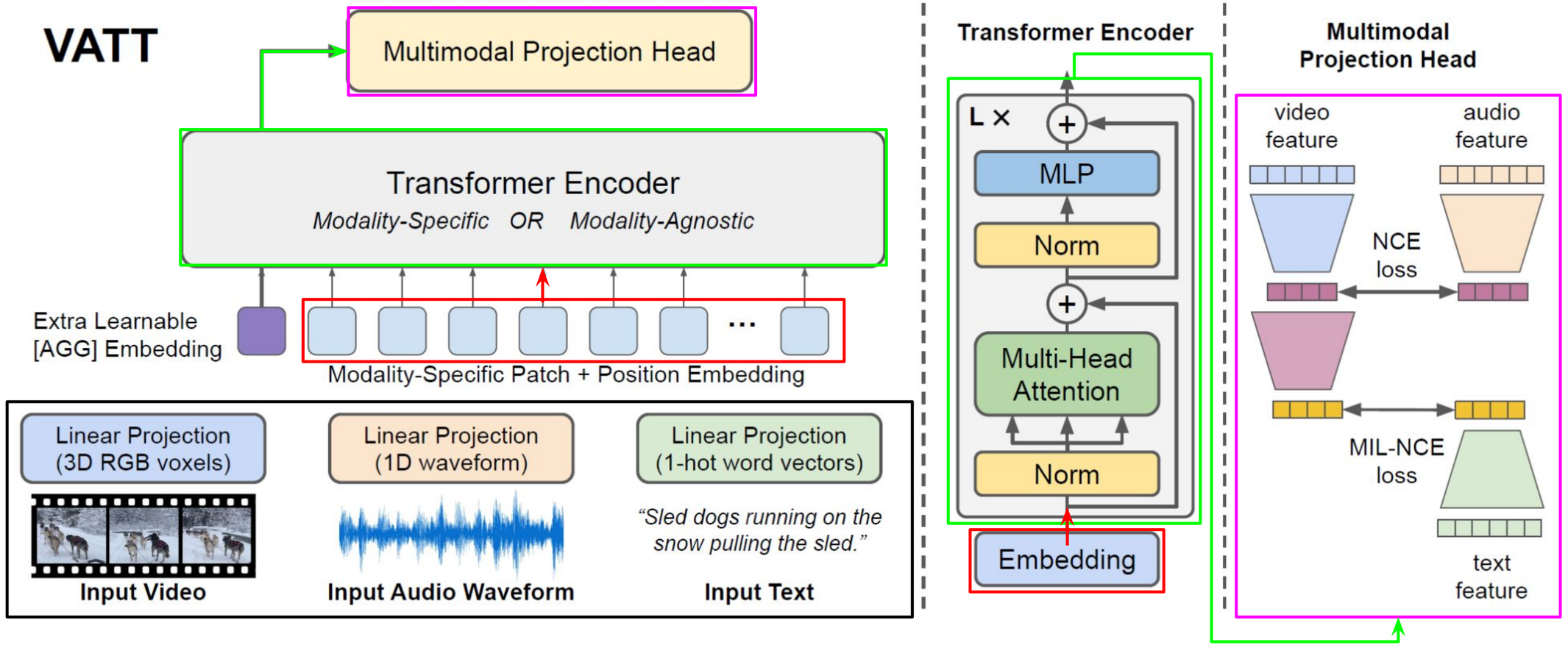
2. Modality-Agnostic

모달리티들에 대해 통합된 **하나의 트랜스포머 인코더를 사용**

1번이 성능은 더 좋지만 2번도 줄어든 연산량에 비해 더 좋은 성능을 보임

#03 Architecture Overview

같은 색은 동일한 내용을 의미



각 모달리티의 트랜스포머 출력값을 공통 공간(Common Space) 으로 매핑해줌
공통 공간에 매핑된 모달리티 쌍의 contrastive 학습 진행

Modality-Specific 방법에서 이어지는 것으로 추측

#03 Multimodal raw data Input

VATT는 인터넷 영상의 raw RGB frame, audio waveform, 음성을 전사한 텍스트를 입력으로 받음

Input Video

영상 프레임의 RGB
3 채널 픽셀값

Input Audio Waveform

음성의 air density
amplitude(진폭) = waveform 형태

waveform이란 파형으로
나타내어지는 음성 데이터
형식을 말함

"Sled dogs running on the snow pulling the sled."

Input Text

단어 시퀀스

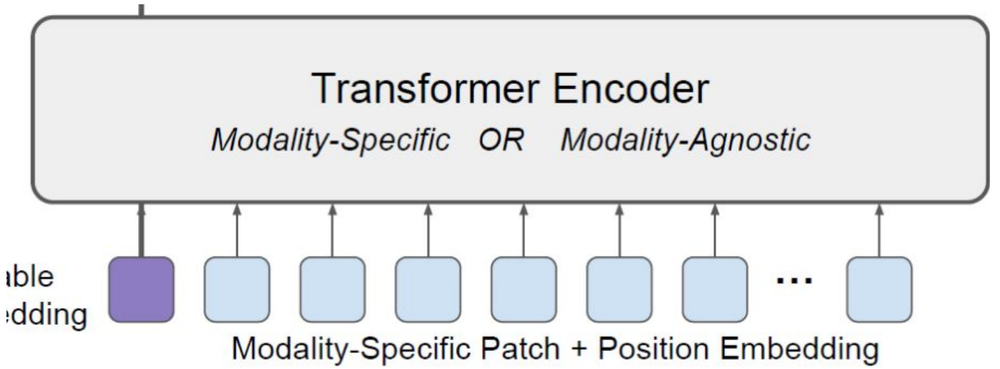
위 세가지 모달이 각각 raw signal (가공되지 않은 데이터)로 입력됨

#03 Tokenization and Positional Encoding

modality-specific tokenization layers

각 모달의 raw signal 형식을 받아 트랜스포머에 입력되기 적합한 벡터 시퀀스로 변환

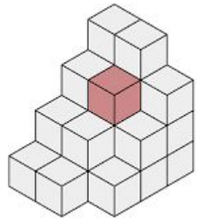
각 모달마다 다른 positional encoding을 적용
→ 각 모달의 law data로 순서화된 토큰을 생성



video to sequence tokenization

1. T x H x W 차원의 전체 비디어 클립을 [T/t] x [H/h] x [W/w] 패치의 시퀀스로 분할 (각 패치는 txhxwx3개의 voxel)

T : temporal, 비디오 프레임 시간대를 의미
H : Height
W : Width



voxel

2D 픽셀을 3차원의 형태(3D)로 구현한 데이터 형식 (ex. 마인크래프트)

$$\mathbf{W}_{vp} \in \mathbb{R}^{t \cdot h \cdot w \cdot 3 \times d}$$

2. 패치에 각 voxel에 (t,h,w,3) 학습 가중치 Wvp를 곱해 d차원 벡터로 linear projection

#03 Tokenization and Positional Encoding

video positional encoding

$$\boldsymbol{e}_{i,j,k} = \boldsymbol{e}_{\text{Temporal}_i} + \boldsymbol{e}_{\text{Horizontal}_j} + \boldsymbol{e}_{\text{Vertical}_k},$$
$$\boldsymbol{E}_{\text{Temporal}} \in \mathbb{R}^{\lceil T/t \rceil \times d}, \quad \boldsymbol{E}_{\text{Horizontal}} \in \mathbb{R}^{\lceil H/h \rceil \times d}, \quad \boldsymbol{E}_{\text{Vertical}} \in \mathbb{R}^{\lceil W/w \rceil \times d}$$

3. [T/t] + [H/h] + [W/w] positional 임베딩으로 각 패치를 인코딩한다

audio sequence tokenization & positional encoding

T'크기의 1차원 데이터로 모델에 입력됨

1. 입력을 [T'/t] 세그멘트로 분할 (각 세그멘트는 t'길이의 waveform amplitude로 이루어짐)
2. 학습 가중치와 곱해서 d차원 벡터로 linear projection

$$\boldsymbol{W}_{ap} \in \mathbb{R}^{t' \times d}$$

3. 비디오와 마찬가지 방법으로 각 waveform 세그멘트들을 [T'/t'] 임베딩

#03 Tokenization and Positional Encoding

text sequence tokenization & positional encoding

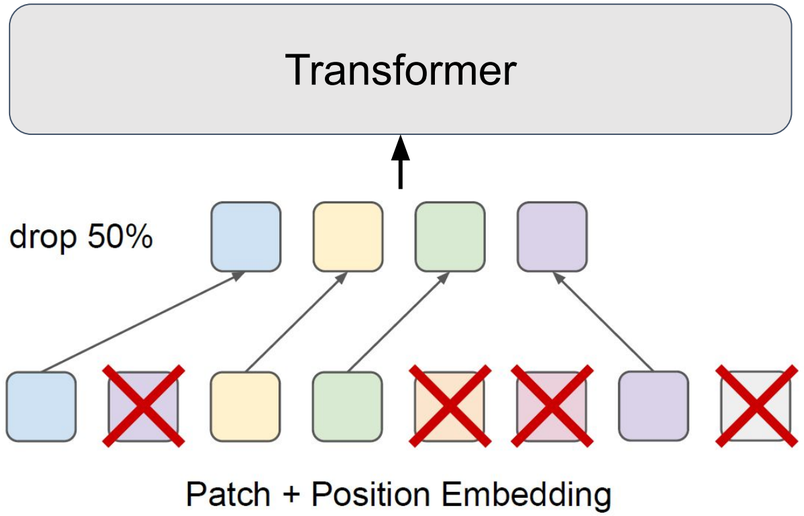
1. 모델이 v길이의 단어를 입력으로 받음
2. 각 단어를 아래 가중치와 곱해서 v 차원의 원-핫 벡터로 linear projection

$$\boldsymbol{W}_{tp} \in \mathbb{R}^{v \times d}.$$

#03 DropToken

VATT 논문에서 모델 복잡도를 낮추기 위해 DropToken 방식을 제안

들어온 입력에서 랜덤하게 토큰을 선택해 트랜스포머 모델에 전달하는 방법으로 Transformer의 계산량을 효과적으로 줄임



Transformer의 계산복잡도

$O(N^2)$ 입력의 길이 N에 대해 계산복잡도는 N의 제곱에 비례

계산복잡도를 낮추기 위해서는 전체 입력이나 트랜스포머에 임베딩되는 입력의 크기를 줄여야 하는데 해당 논문에서는 후자 방식을 제안

#03 DropToken Results

영상 프레임이나 오디오 데이터는 인접한 부분에서 중복성을 가지기 때문에 고해상도 영상의 입력을 그대로 유지한 채 임베딩 단계에서 토큰 개수를 줄이는 것이 더 효율적이라고 가정

영상, 오디오 데이터에서만 droptoken을 적용한 결과

저해상도

고해상도

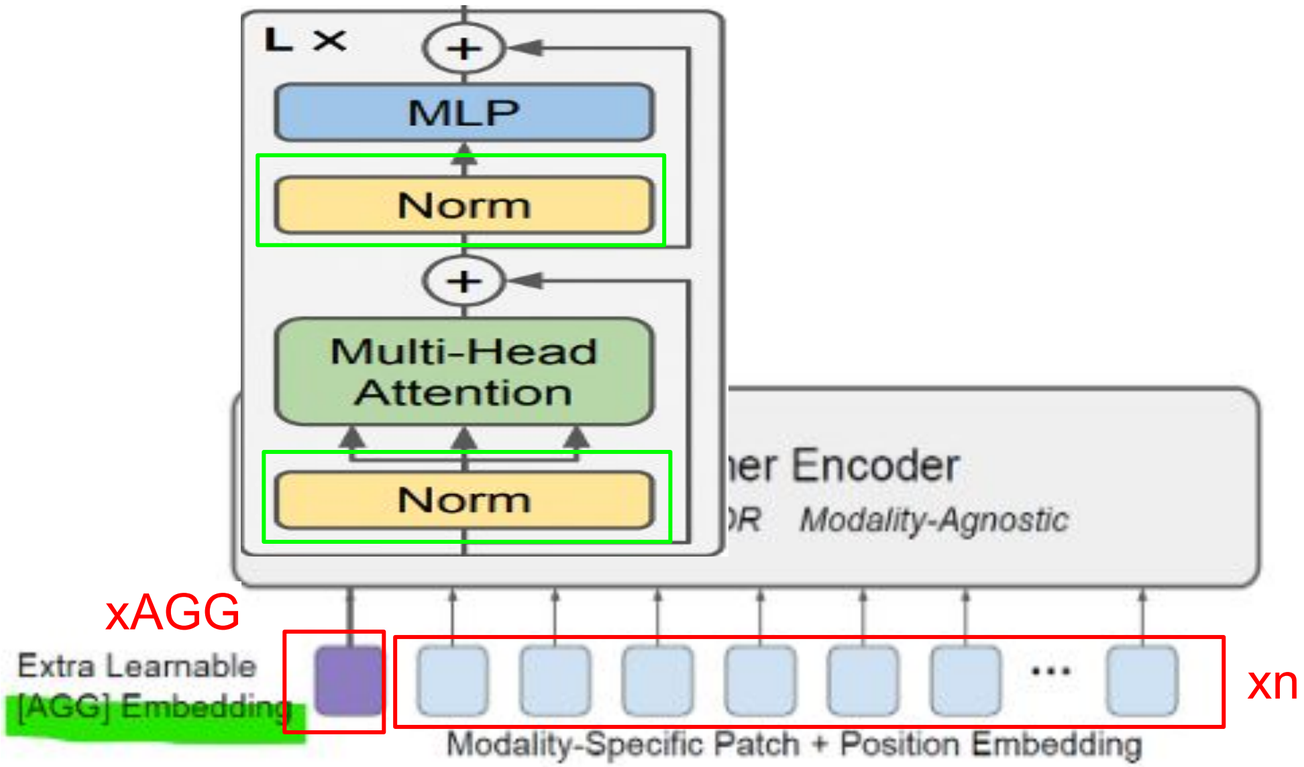
Resolution/ FLOPs	DropToken Drop Rate 75%	50%	25%	0%
32 × 224 × 224	-	-	-	79.9
Inference (GFLOPs)	-	-	-	548.1
64 × 224 × 224	-	-	-	80.8
Inference (GFLOPs)	-	-	-	1222.1
32 × 320 × 320	79.3	80.2	80.7	81.1
Inference (GFLOPs)	279.8	572.5	898.9	1252.3

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

고해상도 이미지에서 DropToken을 50%를 사용했을 때 저해상도에서 DropToken을 사용하지 않았을 때보다 좋은 성능을 보임

#03 Transformer Encoder

0. 기호 설명



$$z_0 = [x_{\text{AGG}}; x_0\boldsymbol{W}_p; x_1\boldsymbol{W}_p; \ldots ; x_N\boldsymbol{W}_p] + e_{\text{POS}}$$

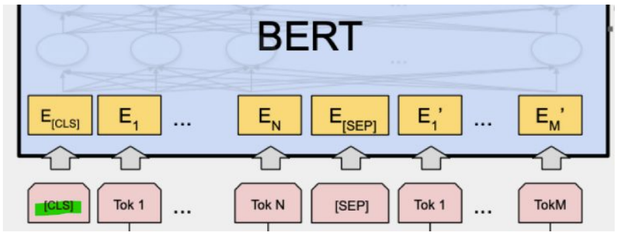
$$z_l' = \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad l = 1 \ldots L$$

$$z_l = \text{MLP}(\text{LN}(z_l')) + z_l', \quad l = 1 \ldots L$$

$$z_{\text{out}} = \text{LN}(z_L)$$

xn : 인풋 패치 시퀀스

xAGG : 입력 시퀀스 전체에 대한 정보들을 합쳐주는 aggregation 토큰으로 BERT모델의 CLS 토큰과 같은 역할을 함



xAGG 토큰을 통해 추후 모달리티들을 common space로 mapping 시켜주거나 분류 테스크를 해결할 수 있다.

트랜스포머 아키텍쳐는 이전의 weight로부터 쉽게 전이학습이 가능하도록 BERT, VIT 구조와 유사하게 설계

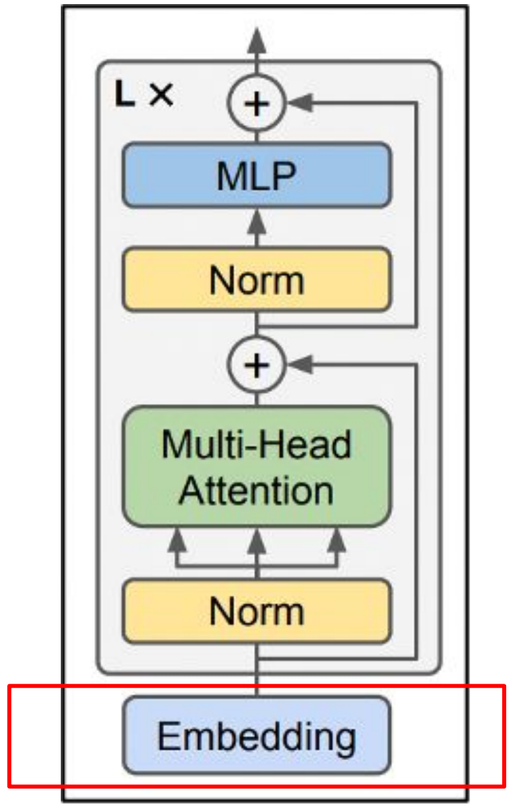
MHA : Multi-Head-Attention

MLP : Multi-Layer Perceptron

LN : Layer Normalization

#03 Transformer Encoder

1. 임베딩 단계



z0

$$z_0 = [x_{\text{AGG}}; x_0\boldsymbol{W}_p; x_1\boldsymbol{W}_p; \ldots ; x_N\boldsymbol{W}_p] + e_{\text{POS}}$$
$$z_l' = \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad l = 1...L$$
$$z_l = \text{MLP}(\text{LN}(z_l')) + z_l', \quad l = 1...L$$
$$z_{\text{out}} = \text{LN}(z_L)$$

각 패치를 가중치와 곱해준 패치 시퀀스가 트랜스포머 모델에 입력됨

ePOS : position encoding

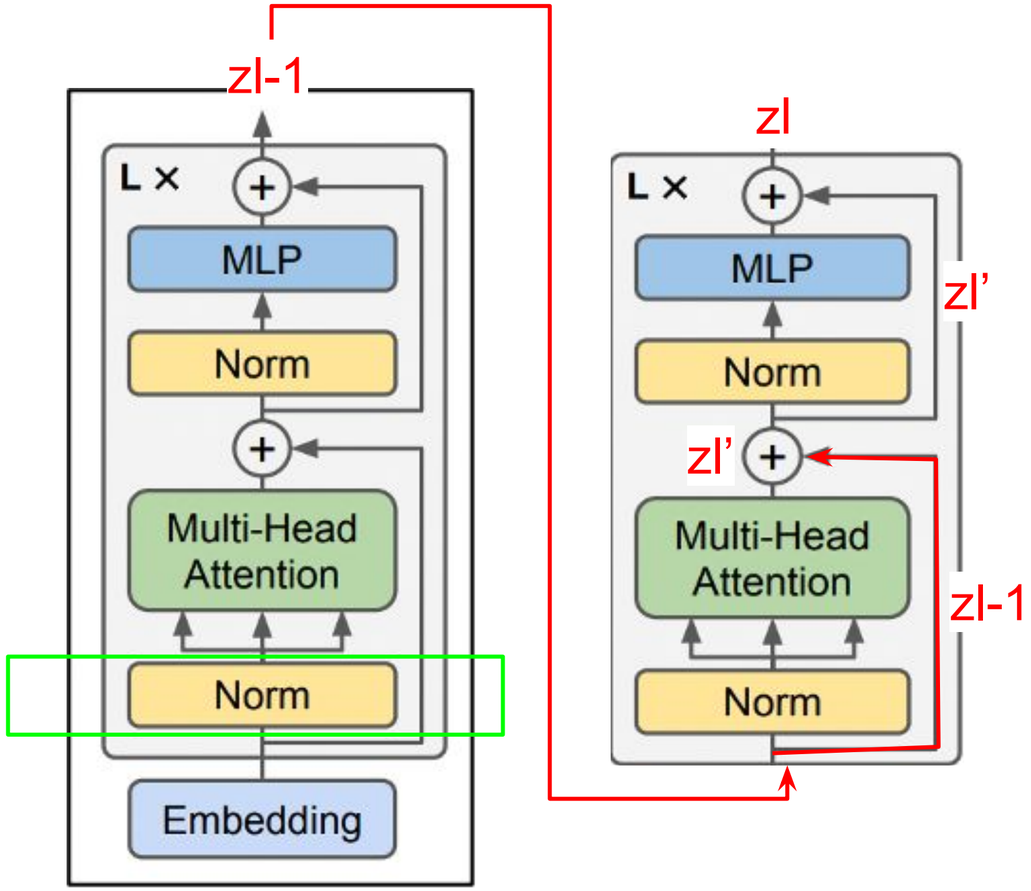
텍스트 모델에서는 position encoding을 제거하는 대신
Multi Head Attention 모듈의 첫 번째 레이어에서 각각의 attention score에 대해 relative bias를 학습하도록 설계

텍스트 모델의 웨이트를 SOTA 언어 모델인 T5에 바로 transfer할 수 있다.

EWHA EURON

#03 Transformer Encoder

2. 트랜스포머



$$z_0 = [x_{\text{AGG}}; x_0\boldsymbol{W}_p; x_1\boldsymbol{W}_p; \ldots ; x_N\boldsymbol{W}_p] + e_{\text{POS}}$$
$$z_l' = \text{MHA}(\text{LN}(z_{l-1})) + z_{l-1}, \quad l = 1 \ldots L$$
$$z_l = \text{MLP}(\text{LN}(z_l')) + z_l', \quad l = 1 \ldots L$$
$$z_{\text{out}} = \text{LN}(z_L)$$

GPT2 모델처럼 layer-normalization을 먼저 수행하고, MHA, MLP 레이어를 통과하도록 설계

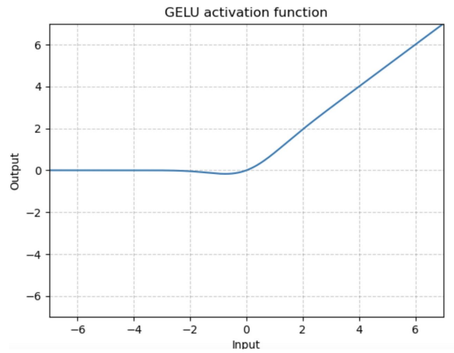
Multi-Head-Attention에서 GeLU를 활성함수로 사용

GeLU (Gaussian Error Linear Unit)

$$\text{GELU}(x) = x * \Phi(x)$$

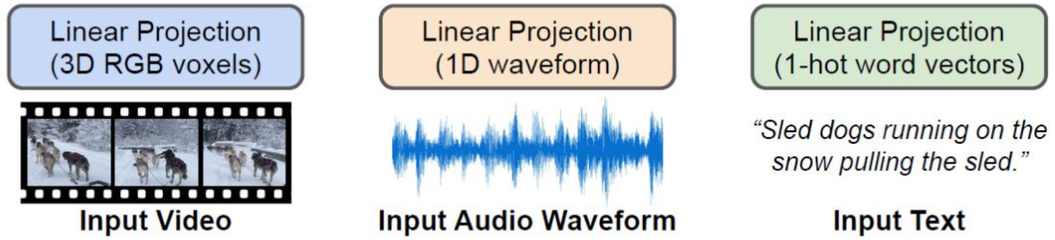
단 Φ(x)는 x의 누적분포함수값

입력의 누적분포함수값을 입력 원본과 곱해줌으로써 전체 분포에서의 상대적 크기를 반영



#03 Common Space Projection

가정 : 서로 다른 모달리티는 서로 다른 수준의 semantic granularity(의미론적 단위)를 가진다.



영상과 음성, 텍스트 데이터의 단위가 모두 달라 모델이
각 모달리티를 비교해서 학습하기 어려움

semantically hierarchical common space mapping

영상-음성, 영상-텍스트 두 단계를 거쳐서
각 모달리티를 같은 길이의 선형 공간으로 매핑시켜
같은 차원에서의 학습이 가능하도록 함

#03 Common Space Projection

1. multi-level projections

영상-음성을 짝지어 512 길이의 공통 벡터 공간으로 linear projection 실행

a. (video - audio) Common Space mapping

$$z_{v,va} = g_{v \to va}(z_{\text{out}}^{\text{video}})$$

$$z_{a,va} = g_{a \to va}(z_{\text{out}}^{\text{audio}})$$

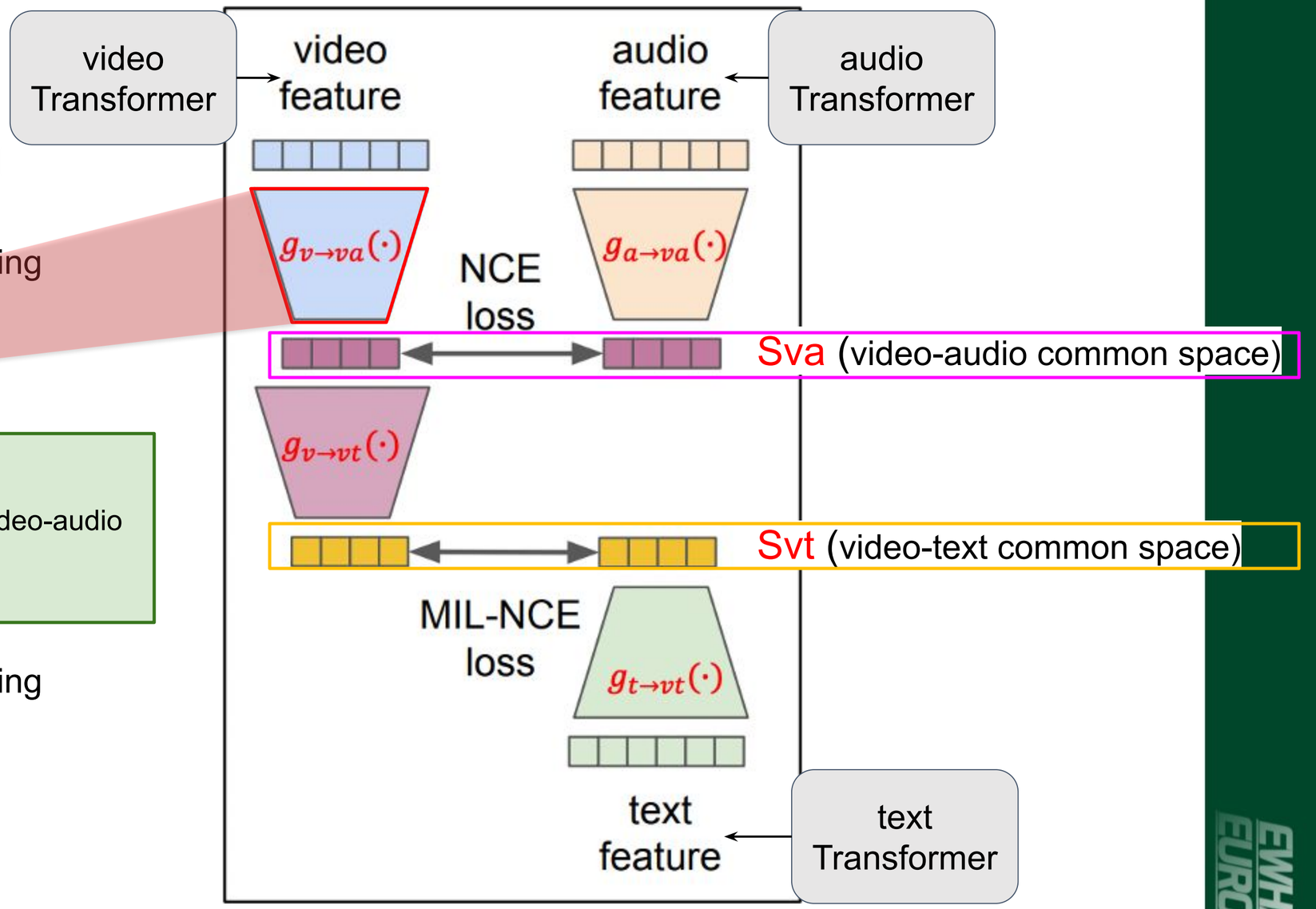
g : projection head

ex. gv->va 는 video transformer의 출력을 받아 video-audio 공통공간으로 **linear projection** 진행
여기서 video-audio common space를 Sva라 부름

b. (video - audio) Common Space mapping

$$z_{t,vt} = g_{t \to vt}(z_{\text{out}}^{\text{text}})$$

$$z_{v,vt} = g_{v \to vt}(z_{v,va})$$



#03 Common Space Projection

1. multi-level projections

영상-음성을 짝지어 512 길이의 공통 벡터 공간으로 linear projection 실행

a. (video - audio) Common Space mapping

$$z_{v,va} = g_{v \to va}(z_{\text{out}}^{\text{video}})$$

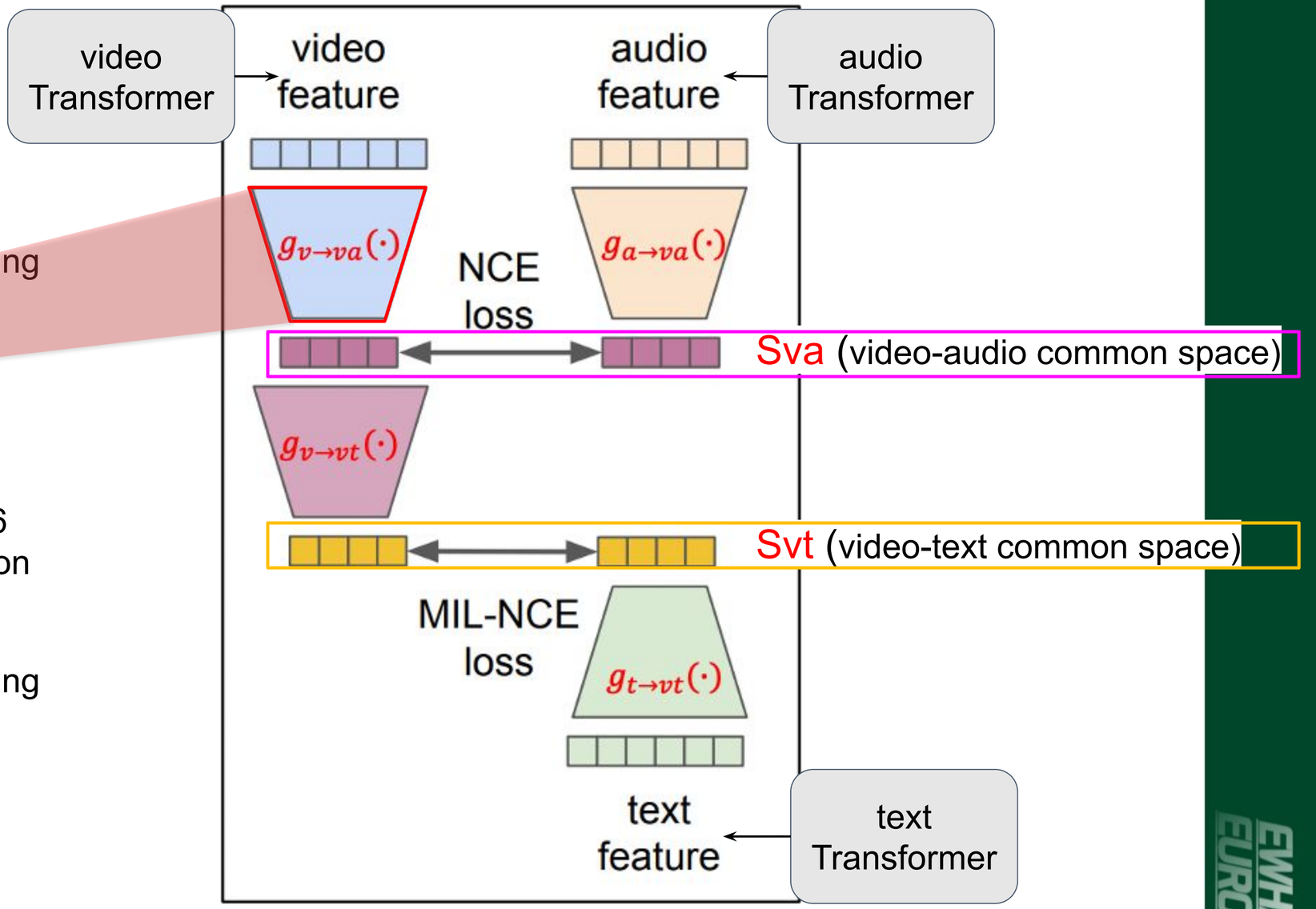
$$z_{a,va} = g_{a \to va}(z_{\text{out}}^{\text{audio}})$$

(Sva에 매핑된) 영상-텍스트를 짝지어 256 길이의 공통 벡터 공간으로 linear projection 실행

b. (video - audio) Common Space mapping

$$z_{t,vt} = g_{t \to vt}(z_{\text{out}}^{\text{text}})$$

$$z_{v,vt} = g_{v \to vt}(z_{v,va})$$



#03 Multimodal Contrastive Learning

비디오-텍스트, 그리고 비디오-오디오 쌍의 자기지도학습에 사용되는 손실 함수로 Noise Contrastive Estimation (NCE)를 적용

positive pair

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

positive pair　　negative pair

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

(video-audio) pairs -> Noise Contrastive Estimation (NCE)
(video-text) pairs -> Multiple Instance Learning NCE (MIL-NCE)

#03 Multimodal Contrastive Learning

Noise Contrastive Estimation (NCE) 등장 배경

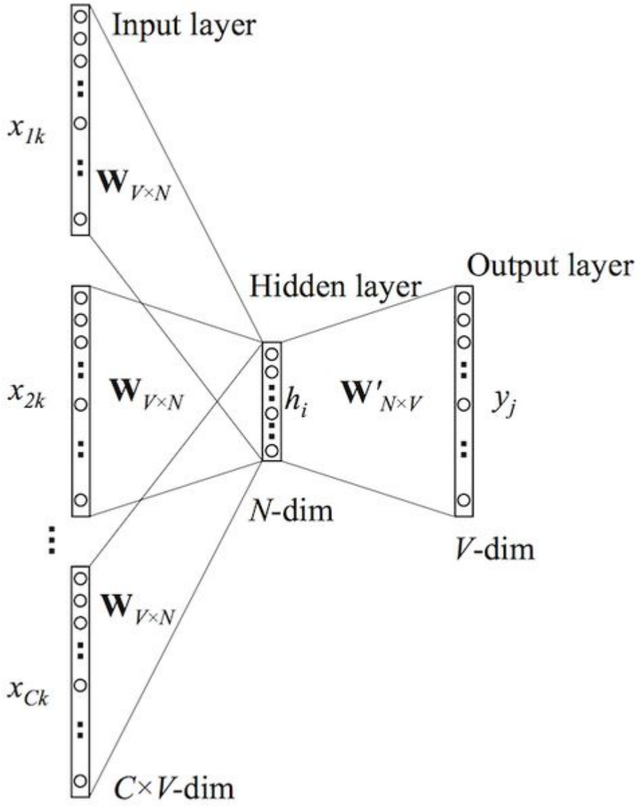
"빠른 주황색 여우가 점프를 한다."

테스크 : 위 문장을 문맥, 타겟으로
분리해 타겟 단어를 맞추고자 함

((빠른, 여우가), 주황색),

((주황색, 점프를), 여우가),

((여우가, 한다), 점프를)



CrossEntropyLoss L의 계산식

y : output이 target 단어들의 후보군에 대한 확률 map

p : softmax 확률값

$$\text{for } i: \text{ index of word, } z_i = Wx_i$$

$$p(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{|V|} e^{z_j}} \qquad L = -\sum_{j}^{|V|} y_j \log(p_j) = -\log(p_{\text{target}})$$

모든 가능한 예측에 대해 손실을 계산하는 기존의 손실 함수는 오직 정답 단어인 경우에만 관심을 가지는 자연어 처리 테스크에 부적절함

p(zi)의 분모가 항상 같은 값을 가지기 때문에 타겟 단어가 아닌 상관없는 단어도 0이 아닌 기울기 값 (즉 **노이즈**)를 가지기 때문, 입력 단어가 길어질수록 더 noisy한 학습 진행

#03 Multimodal Contrastive Learning

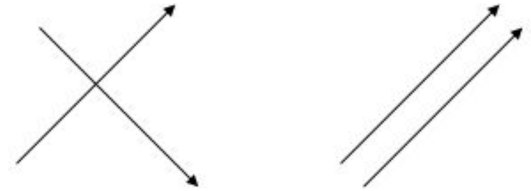
Noise Contrastive Estimation (NCE)

NCE 손실함수는 이를 해결하기 위해 positive pair(타겟)을 1, 그리고 negative pair을 0으로 지표화한다. 이로써 모델은 데이터 내의 positive, negative 샘플 관의 관계를 학습하고 타겟의 확률값에만 집중하게된다.

모델 적용 방법

video-audio, video-tex 쌍을 다른 시간대로부터 가져온다고 가정할 때

positive pair : 두 modality가 같은 video clip에서 선택된 경우
negative pair : 두 modality가 다른 video clip에서 선택된 경우



각 쌍의 유사도를 cosine similarity로 계산

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

positive pair (분자) / positive pair (분모) / negative pair

NCE loss를 최소화하기 위해서는 positive pairs의 분자와 positive + negative pairs가 더해진 분모의 크기가 같아져야 하므로 모델은 positive pairs 지표는 1에, negative pairs 지표는 0에 가까워지도록 학습하게됨

→ **NCE 목적함수는 positive pair간 유사도를 최대화하고 negative pair간 유사도를 최소화**

#03 Multimodal Contrastive Learning

Multiple Instance Learning NCE (MIL-NCE)

text 데이터셋은 규격화된 ASR(자동 자막 생성 api)를 사용하고 video는 음성이나 자막이 없는 경우도 많기 때문에 노이즈가 많음, 따라서 NCE의 확장 버전인 ML-NCE 사용

영상과 시간적으로 인접한 여러개의 text를 모두 positive pair로 가정 -> positive pair 셋이 하나가 아닌 여러개

P(z) : positive pair, 구체적으로 video clip과 시간적으로 가장 가까운 5개의 text clip
N(z) : negative pair (그 외)
τ : negative pair로부터 positive pair 구분의 부드러운 정도를 조절하는 변수

positive pair

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

positive pair　　negative pair

VATT의 전체 목적함수

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

λ는 2개의 loss의 비율을 조절

EWHA EURON

Experiments

EWHA EURON

#04 Experimental Setup

Pre-Training Datasets

AudioSet dataset

2백만

video-audio
(text 제외)

유튜브 영상을 10초 클립 단위로 자른 AudioSet 데이터셋의 video-audio 데이터만을 사용

HowTo100M dataset

136M

video-audio-text

유튜브 나레이션 영상 클립 + 자동생성 자막 데이터

유튜브에서 "How to ~"로 시작하는 비디오를 수집해 23k개 도메인에 대한 설명 영상 데이터셋을 만듬

(장점 : 각 장면 관의 명확한 인과관계, 시각 정보와 텍스트 정보의 높은 연관성)

Training time

Medium-Base-Small (MBS; 264M) modality specific VATT 모델 기준
batch size 2048, 256개의 TPUs (v3)으로 pre-training을 하였을 때 3일 정도의 시간이 소요됨

EWHA EURON

#04 Experimental Setup

Downstream Tasks

1. Video action recognition
2. Audio event classification
3. Zero-shot video retrieval

   video-text 간 공통공간 representation의 품질 평가
4. Image classification

   이미지와 영상 간 도메인 차이가 존재하긴 하지만 ImageNet 분류 테스크로 파인튜닝하여 모델의 전이학습 가능성을 평가 (이미지를 영상과 같은 형식으로 입력으로 주기 위해 이미지를 4번 복사하여 네트워크에 집어넣는 방식을 사용)

VATT network sizes

modality-agnostic

- Medium model (VATT-MA; 155M parameters)

modality-specific

- Base-Base-Small (BBS; 197M)
- Medium-Base-Small (MBS; 264M)
- Large-Base-Small (LBS; 415M)

#04 Results

1. Fine-tuning for video action recognition

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

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

Fine Tuning dataset

human action 비디오 클립 데이터셋

model specific 방식의 VATT-Large 모델에서 SOTA 성능을 보여줌

modality-specific

modality-agnostic

#04 Results

1. Fine-tuning for video action recognition

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

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

Fine Tuning dataset

human action 비디오 클립 데이터셋

model agnositc 방법의 경우에도 VATT-Base와 비슷한 성능을 보임

-> 단순히 하나의 트랜스포머 모델만을 사용해도 좋은 성능을 나타낼 수 있음을 암시

modality-specific

modality-agnostic

#04 Results

2. Fine-tuning for audio event classification

METHOD	mAP	AUC	d-prime
DaiNet [25]	29.5	95.8	2.437
LeeNet11 [59]	26.6	95.3	2.371
LeeNet24 [59]	33.6	96.3	2.525
Res1dNet31 [52]	36.5	95.8	2.444
Res1dNet51 [52]	35.5	94.8	2.295
Wavegram-CNN [52]	38.9	96.8	2.612
VATT-Base	**39.4**	**97.1**	**2.895**
VATT-MA-Medium	39.3	97.0	2.884

Table 5. Results for audio event classification on AudioSet.

Modality-Specific, Modality-agnostic 두 모델 모두 CNN 기반 모델보다 일관되게 좋은 성능을 보여줌

Modality-agnostic 모델도 VATT-Base에 비해서 나쁘지 않은 성능을 보이는 것을 확인

#04 Results

3. Fine-tuning for image classification

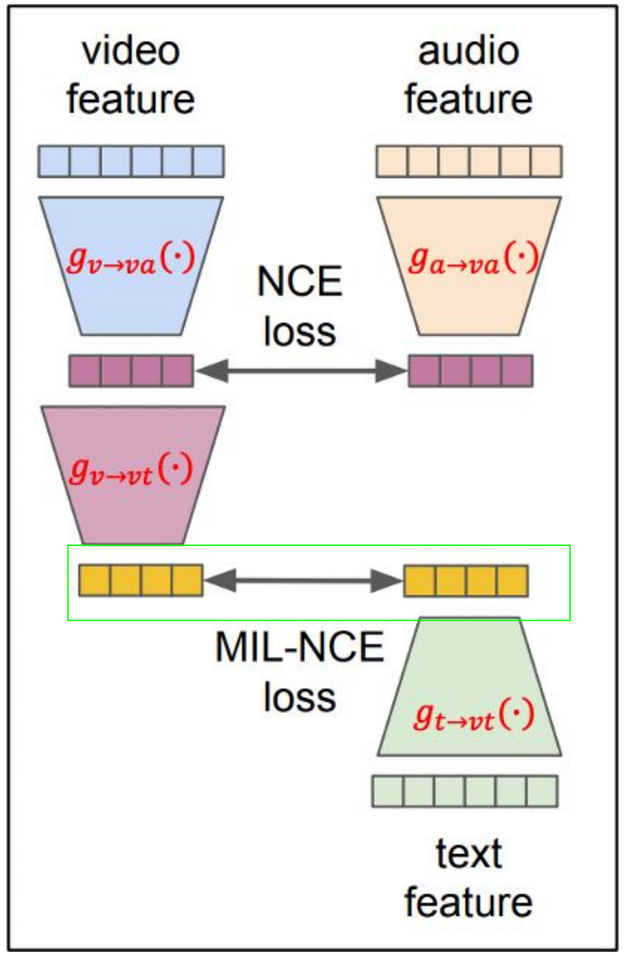
METHOD	PRE-TRAINING DATA	TOP-1	TOP-5
iGPT [18]	ImageNet	66.5	-
ViT-Base [29]	JFT	**79.9**	-
VATT-Base	-	64.7	83.9
VATT-Base	HowTo100M	78.7	93.9

Table 6. Finetuning results for ImageNet classification.

VATT-Base 모델이 이미지 영역에서도 훌륭한 성능을 보임을 확인

#04 Results

4. Zero-shot retrieval



각 모델에 대해 가장 효과적인 Batch 사이즈 실험

video-caption pair 데이터셋

METHOD	BATCH	EPOCH	YouCook2 R@10	YouCook2 MedR	MSR-VTT R@10	MSR-VTT MedR
MIL-NCE [64]	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

video-caption(text) 짝지어진 데이터셋으로 파인튜닝하여 모델이 video와 text를 잘 매핑할 수 있는지를 검증하고자 함
주어진 Text query에 대해 파인튜닝 데이터셋의 비디오들을 텍스트와의 유사성 순으로 정렬

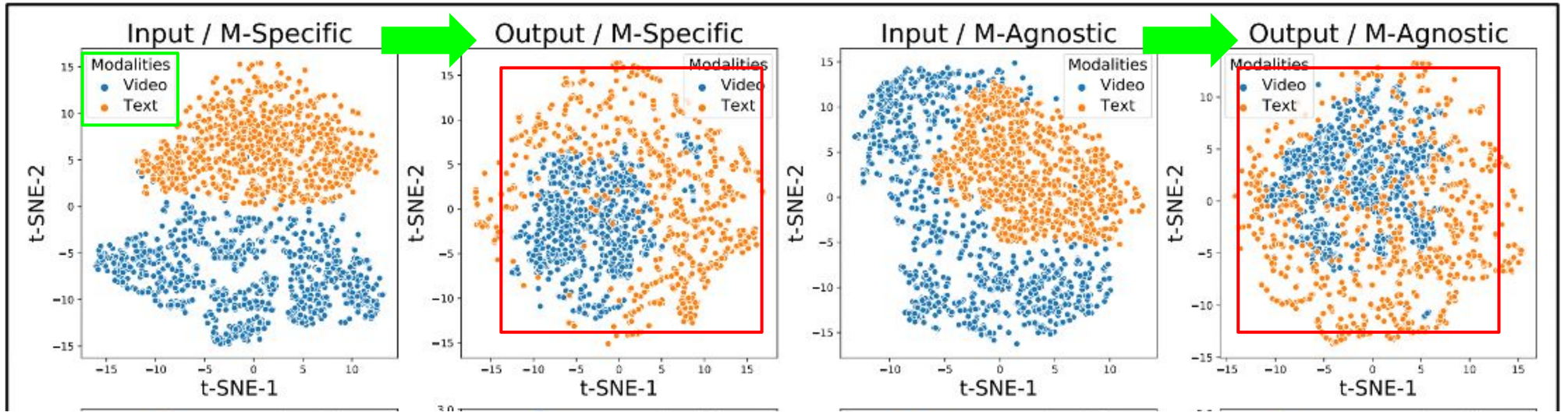
R@10 : 유사도가 높은 top 10 비디오들의 recall (유사도를 recall해준다고 이해함)
MedR : 정렬된 rank의 중앙값

텍스트가 noisy하기 때문에 VATT와 같은 복잡한 언어모델이 낮은 평가를 받는 것으로 예상

EWHA EURON

#04 Feature visualization

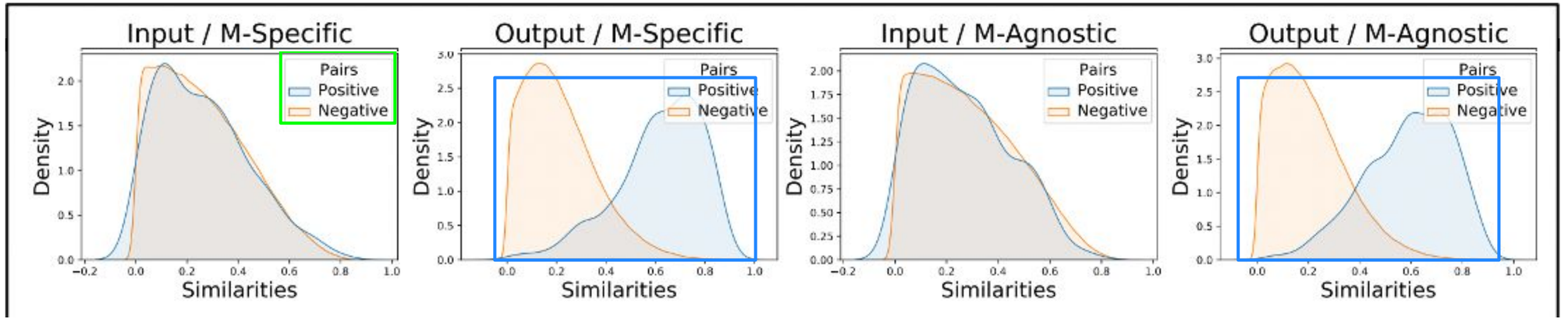
트랜스포머 통과 전, 후 모달리티들의 분포를 t-SNE를 통해 2차원 평면에 시각화한 결과



트랜스포머 통과 후 modality들이 서로 비교적 잘 어우러짐을 확인

VATT 모델이 서로 다른 modality들의 정보들을 잘 표현되도록 학습한다

트랜스포머 video-text쌍의 pair-wise 유사도를 시각화한 결과 positive, negative를 잘 구분해냄



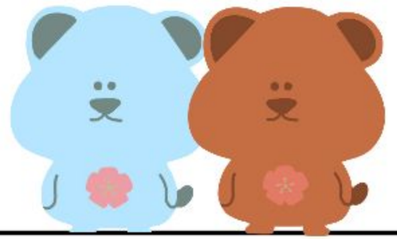
Conclusion

EWHA EURON

#05 결론

- Multimodal 영상 입력에서 작동하는 순수 attention 기반 모델을 구현
- 대규모 자기지도 사전학습 제안으로 비전 분야에서 Transformer 구조의 데이터 부족 문제를 해결
- 여러 downstream task에서 CNN과 비교해 성능이 좋다
- DropToken 방식으로 Transformer의 연산량을 획기적으로 줄여 계산량 대비 성능을 끌어올림

QNA 및 의견



EWHA EURON