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

Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, *34*, 24206-24221.

육현준

1. Introduction

1.1 Background

- Transformer[1]의 등장 이후 자연어 처리(NLP)에서 SOTA 달성
  - High Computational efficiency & Scalability
  - GPT, BERT
- Computer Vision
  - Large-scale Supervised Pre-trained Transformer (ViT[2])의 성공
- Video Recognition task로 확장

1)Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
2)Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

1. Introduction

1.2 Problem

- 대량의 data → <u>Supervised</u> Pretrain
  - 많은 양의 Parameter와 Hyperparameter
  - Bias → 더 많은 양의 labeled data 필요

1.3 Difficulties

- 충분한 양의 label data를 위한 비용과 학습 시간↑
  - Computer Vision에서 Transformer의 적용이 어렵다.

1. Introduction

1.4 Solution

- Unlabeled data
- Raw signal을 입력으로 받는 Self-Supervised Learning Transformer

2. Related Work

2.1 Transformer in Vision

- (ViT) 대량의 label data → Pre-train
  - CNN-base 모델보다 높은 성능
  - 다양한 downsteam task에 활용

2. Related Work

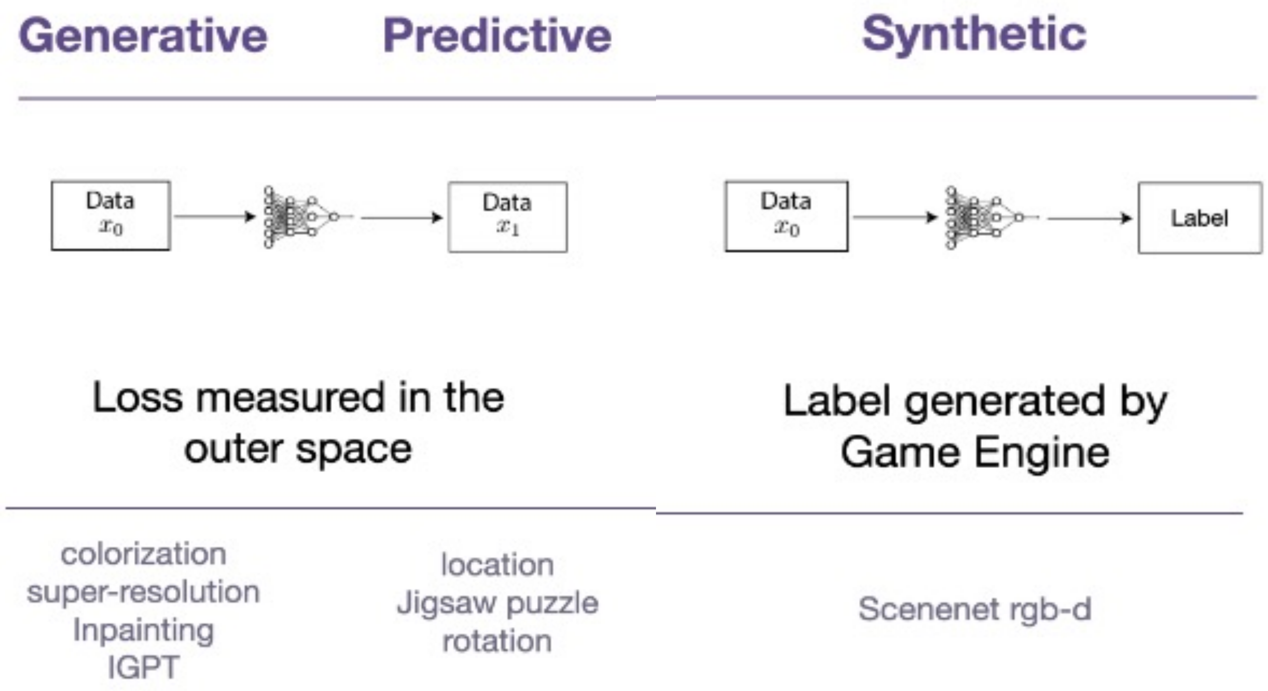
2.2 Self-Supervised Learning

- Single vision modality
  - Self-supervised visual representation learning
    - Pretext task → Contrastive learning

2. Related Work

2.2 Self-Supervised Learning

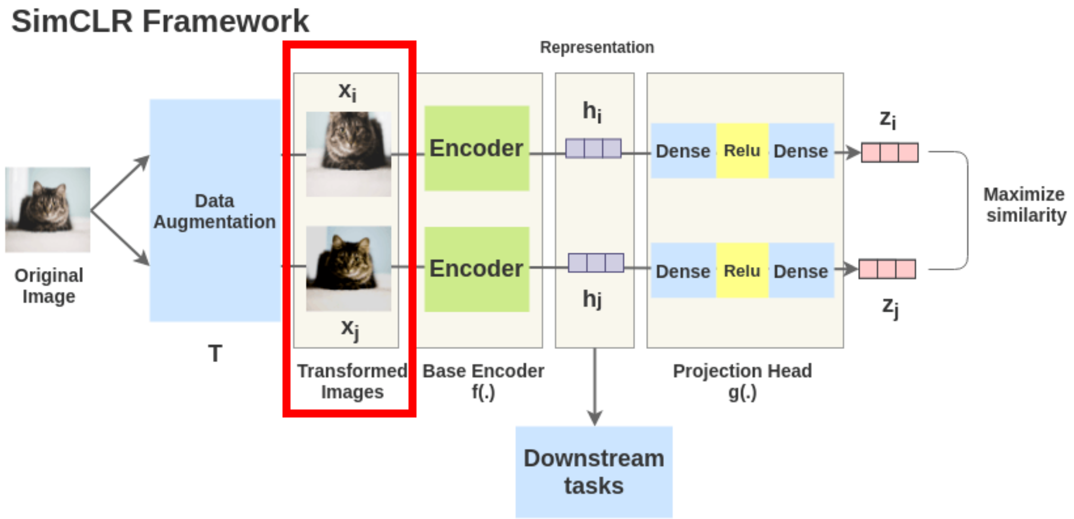
- Single vision modality
  - Pretext task
    - 사람이 정의한 작업을 통해 unlabel data로부터 feature를 추출



2. Related Work

2.2 Self-Supervised Learning

- Single vision modality
  - Contrastive Learning
    - 입력 sample 간의 비교를 통해 representation 학습



3) Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

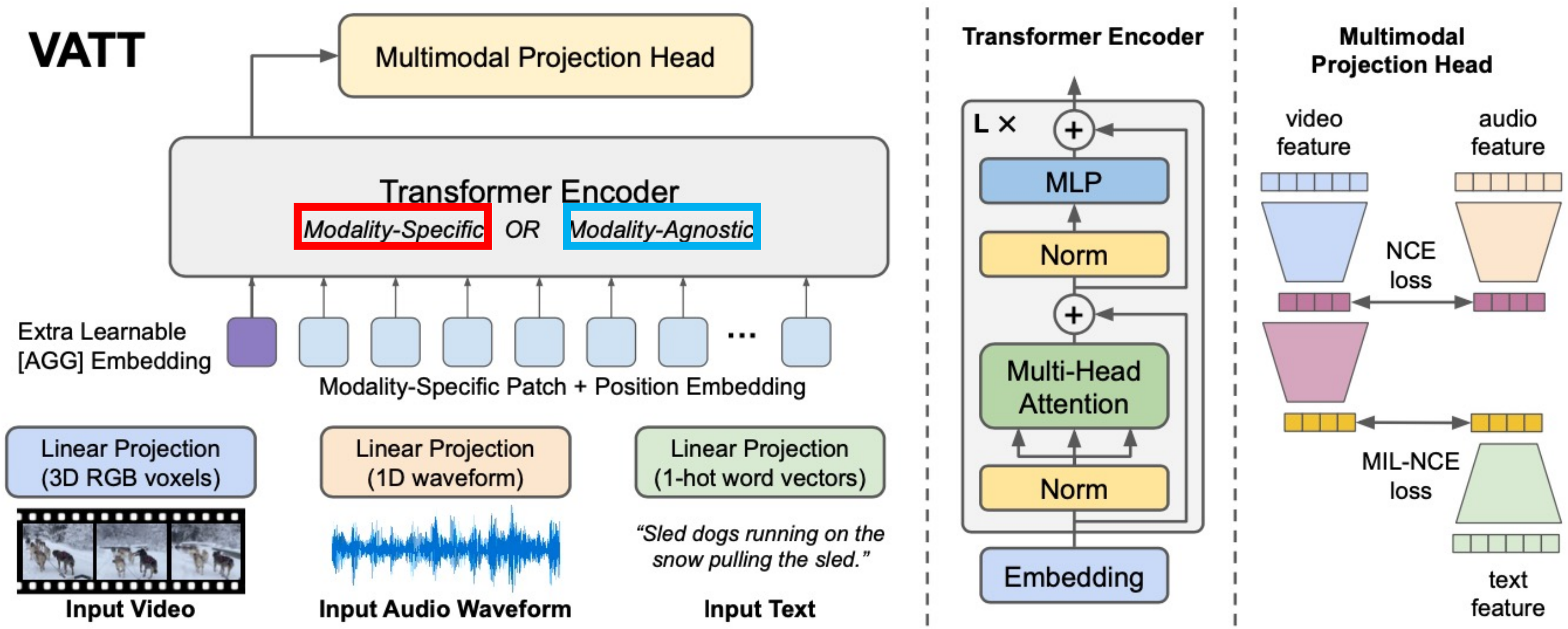
2. Related Work

2.2 Self-Supervised Learning

- Multimodal Video
  - Audio waveform
  - Text scripts
  - Video frames

3. Proposed Idea

3.1 VATT

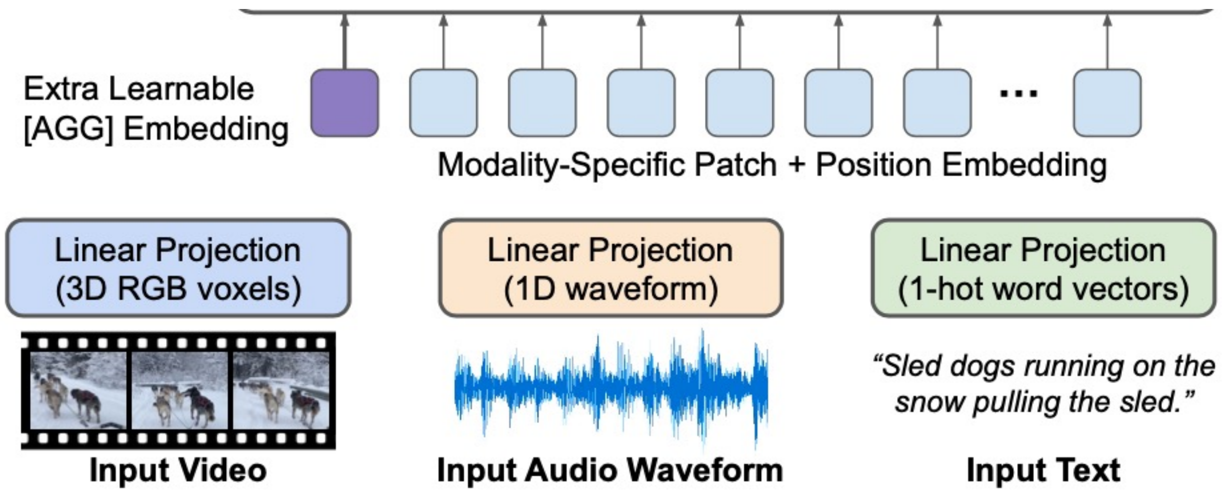


- Transformer : BERT, ViT

- Modality – Specific
- Modality – Agnostic

3. Proposed Idea

3.2 Tokenization and Positional Encoding

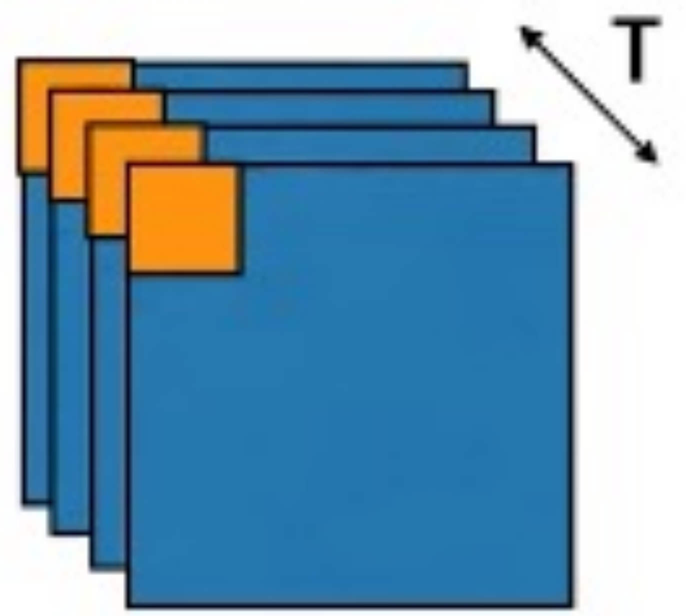


- Raw signal을 input

3. Proposed Idea

3.2 Tokenization and Positional Encoding

- Video
  - 전체 $T \times H \times W \times 3 \ \rightarrow t \times h \times w\ 3$
    - [T/t] x [H/h] x [W/w] patches
  - D-dimension projection (flatten & linear)
    - $W_{vp} \in \ \mathbb{R}^{t \cdot h \cdot w \cdot 3 \times d} \rightarrow$ Transformer input



3. Proposed Idea

3.2 Tokenization and Positional Encoding

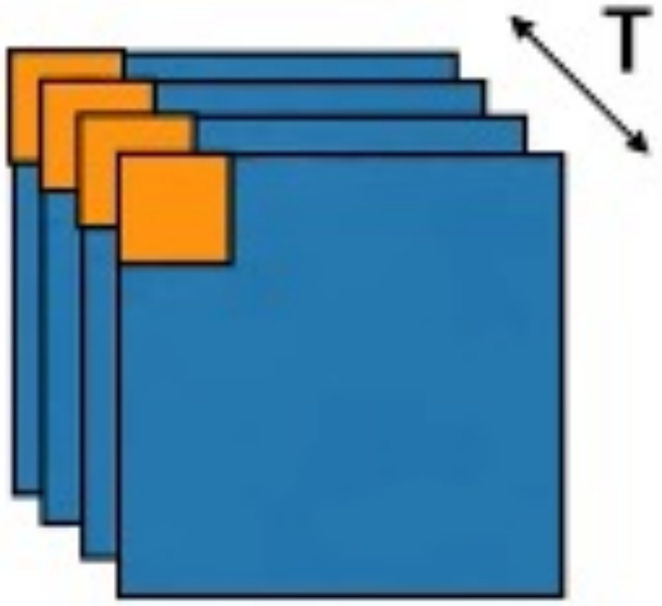
- Positional Encoding
  - Learnable parameter

$$\boldsymbol{E}_{\text{Temporal}} \in \mathbb{R}^{\lceil T/t \rceil \times d}$$

$$\boldsymbol{E}_{\text{Horizontal}} \in \mathbb{R}^{\lceil H/h \rceil \times d}$$

$$\boldsymbol{E}_{\text{Vertical}} \in \mathbb{R}^{\lceil W/w \rceil \times d}$$

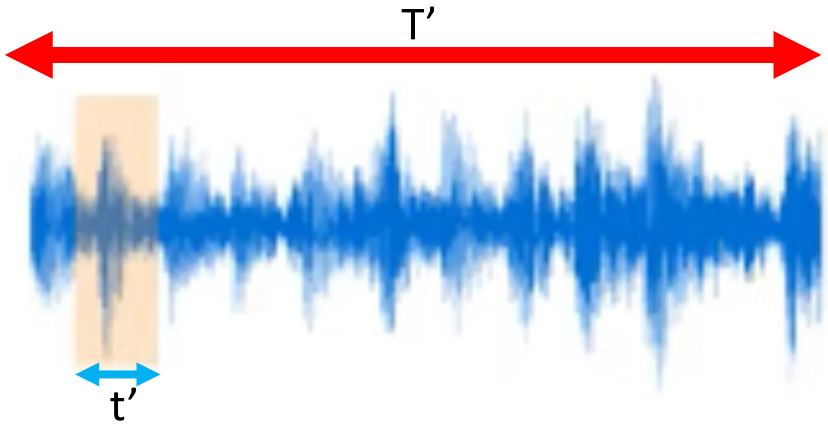
$$\boldsymbol{e}_{i,j,k} = \boldsymbol{e}_{\text{Temporal}_i} + \boldsymbol{e}_{\text{Horizontal}_j} + \boldsymbol{e}_{\text{Vertical}_k}$$



3. Proposed Idea

3.2 Tokenization and Positional Encoding

- Audio
  - 전체 T’ → t'
    - [T’/t’] segments
  - D-dimension projection (linear)
    - $W_{ap} \in \mathbb{R}^{t' \times d}$ → Transformer input



3. Proposed Idea

3.2 Tokenization and Positional Encoding

- Positional Encoding
  - Learnable embedding
    - [T’/t’] 개의 learnable embedding

3. Proposed Idea

3.2 Tokenization and Positional Encoding

- Text
  - V dimension one-hot encoding mapping
  - D-dimension projection (linear)
    - $W_{tp} \in \mathbb{R}^{v \times d} \rightarrow$ Transformer input

```
['나', '는', '자연어', '처리', '를', '배운다']
```

```
단어 집합 : {'나': 0, '는': 1, '자연어': 2, '처리': 3, '를': 4, '배운다': 5}
```

```
one_hot_encoding("자연어", word_to_index)
```

```
[0, 0, 1, 0, 0, 0]
```

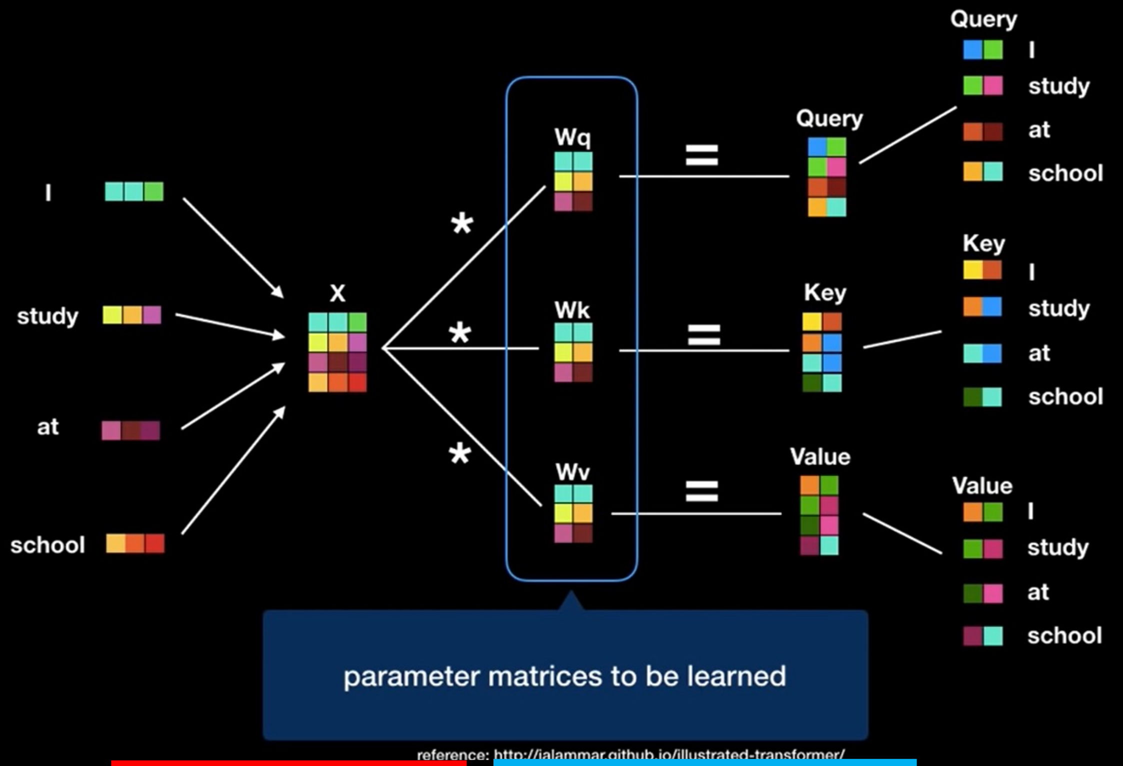
3. Proposed Idea

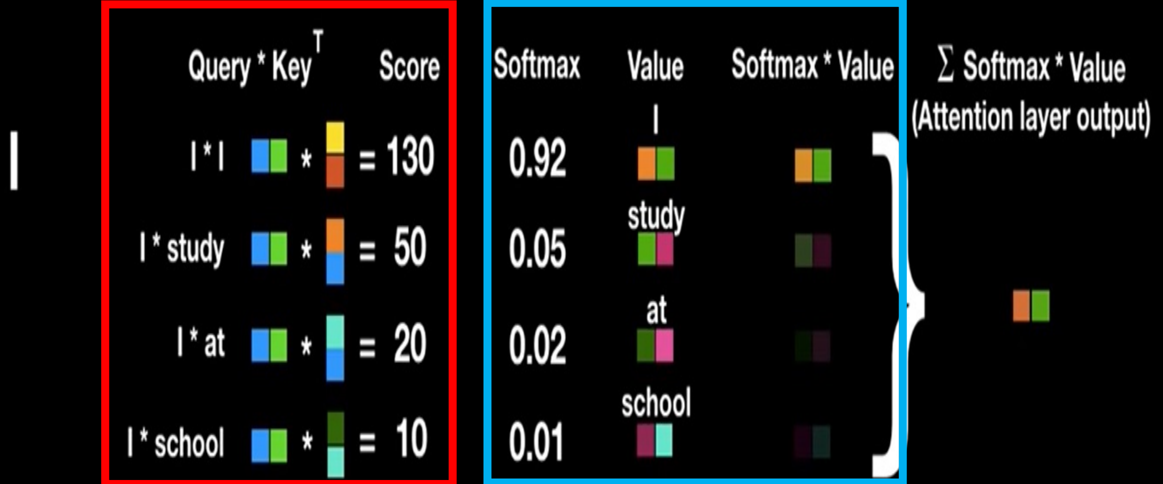
3.2 Tokenization and Positional Encoding

- Relative positional encoding
  - T5[4] model에서 사용한 방법
  - Learnable parameter
  - Attention score + relative bias
- 이 방법을 사용하여 T5모델 Transfer 가능

4) Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

Self Attention





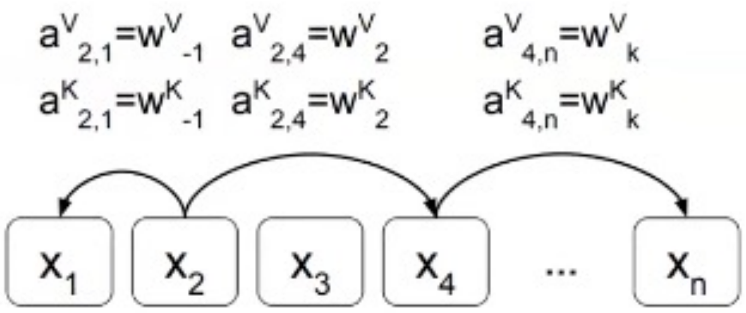
$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V)$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}}$$

$$\downarrow$$

$$e_{ij} = \frac{x_i W^Q (x_j W^K + \boxed{a_{ij}^K})^T}{\sqrt{d_z}}$$

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + \boxed{a_{ij}^V})$$



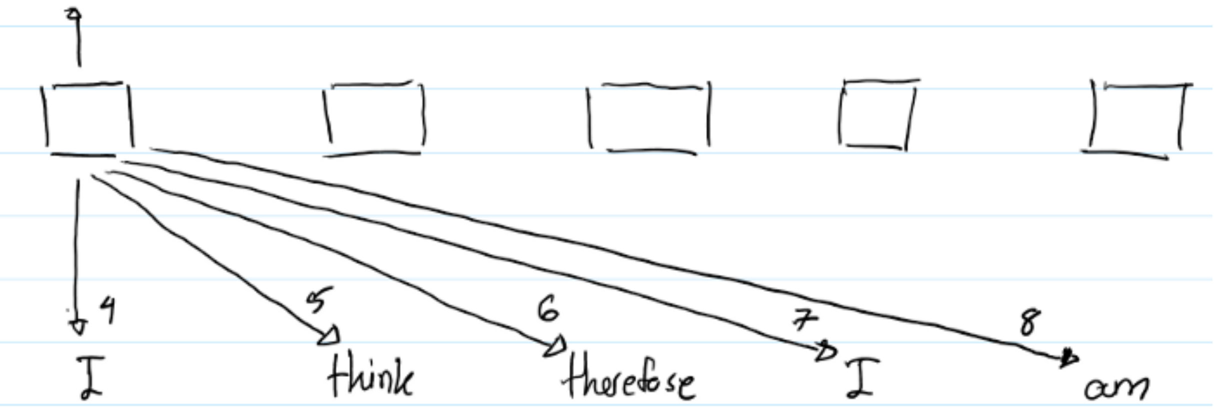
$$a_{ij}^K = w^K_{\mathrm{clip}(j-i,k)}$$

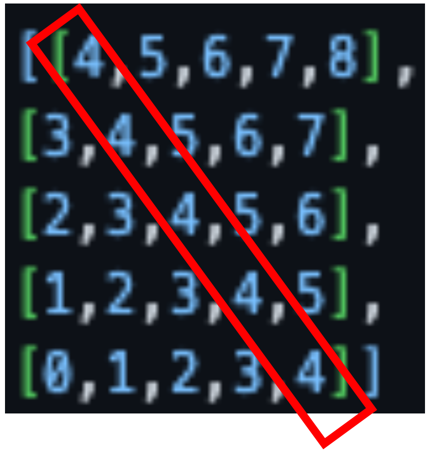
$$a_{ij}^V = w^V_{\mathrm{clip}(j-i,k)}$$

$$\mathrm{clip}(x,k) = \max(-k, \min(k,x))$$

- I think therefore I am (k=4)

Index	Interpretation
0	dist between word at position i and i-4
1	dist between word at position i and i-3
2	dist between word at position i and i-2
3	dist between word at position i and i-1
4	dist between word at position i and i
5	dist between word at position i and i+1
6	dist between word at position i and i+2
7	dist between word at position i and i+3
8	dist between word at position i and i+4

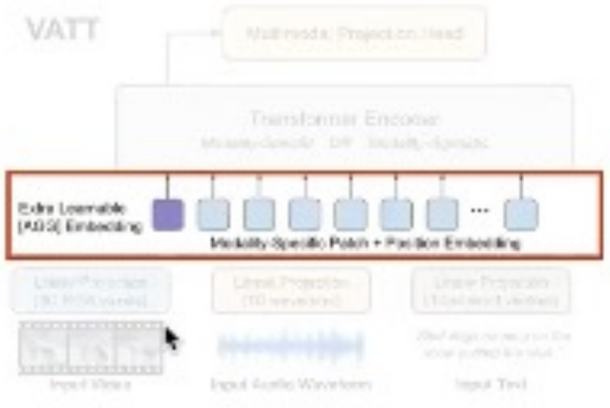




3. Proposed Idea

3.3 DropToken

- $O(N^2)$: Computational Complexity ↓
  - N : 입력 시퀀스의 토큰 수
- 제한된 하드웨어에서 대형 모델을 호스팅 가능
- resolution, dimension을 줄이는 것보다 더 나은 방법
- Video와 audio token에 적용
- 특히, Video는 중복성↑ 효율↑



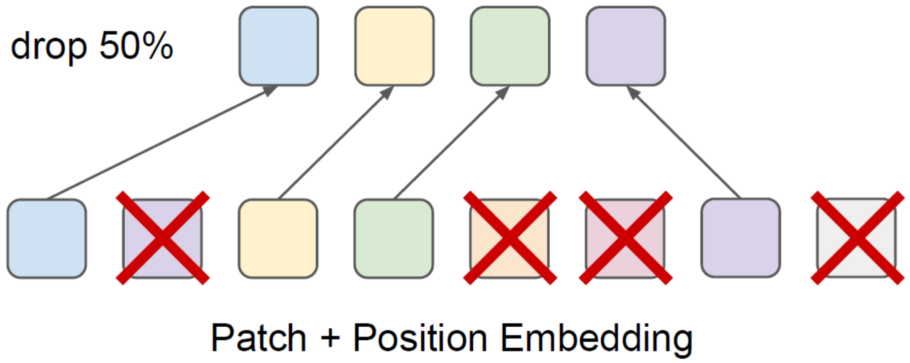
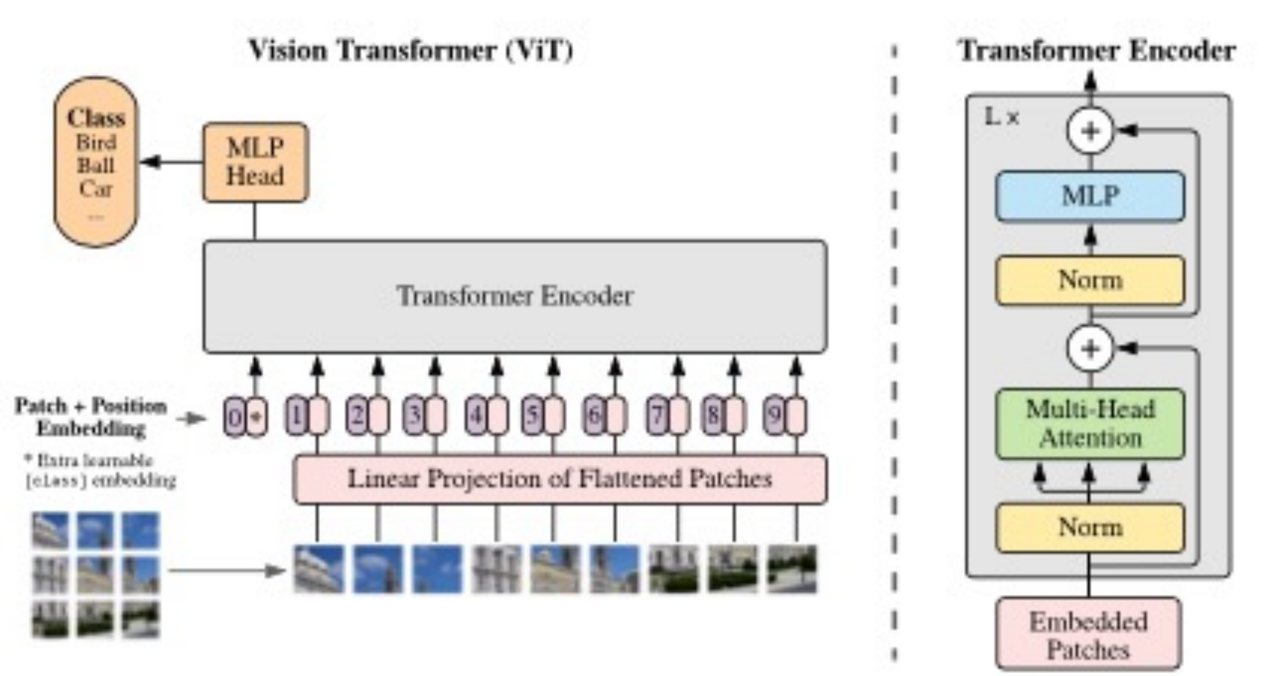


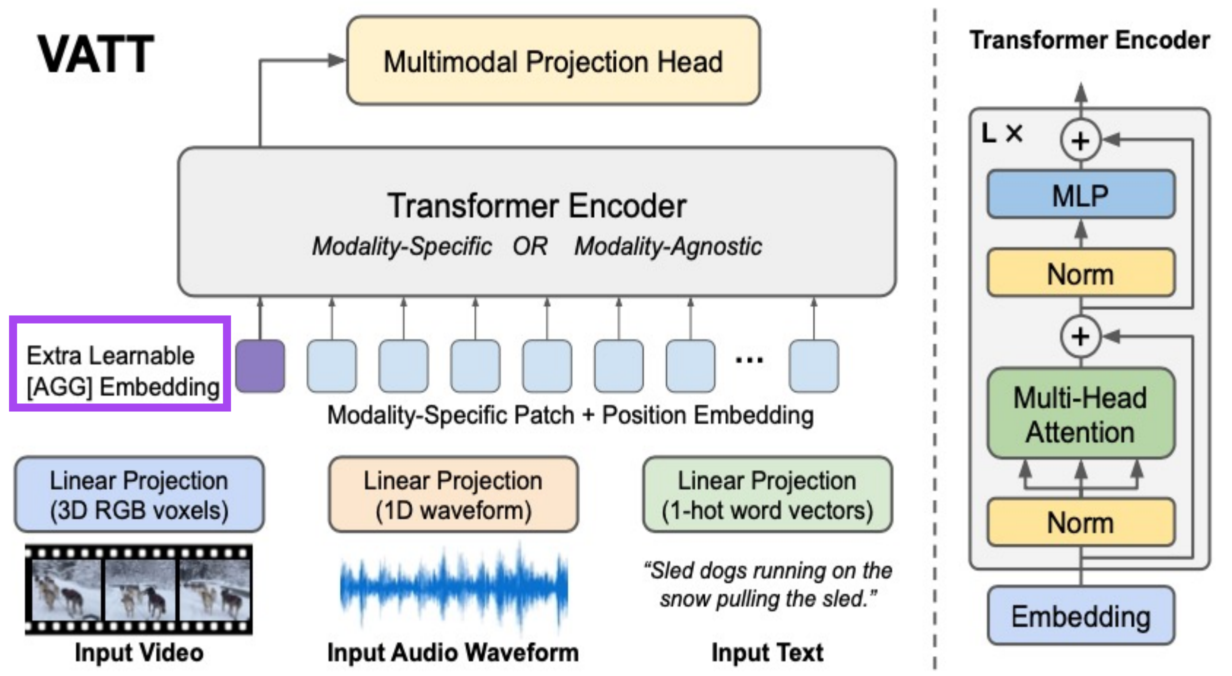
Figure 2. **DropToken**. During training, we leverage the high redundancy in multimodal video data and propose to randomly drop input tokens. This simple and effective technique significantly reduces training time with little loss of quality.

3. Proposed Idea

3.4 The Transformer Architecture

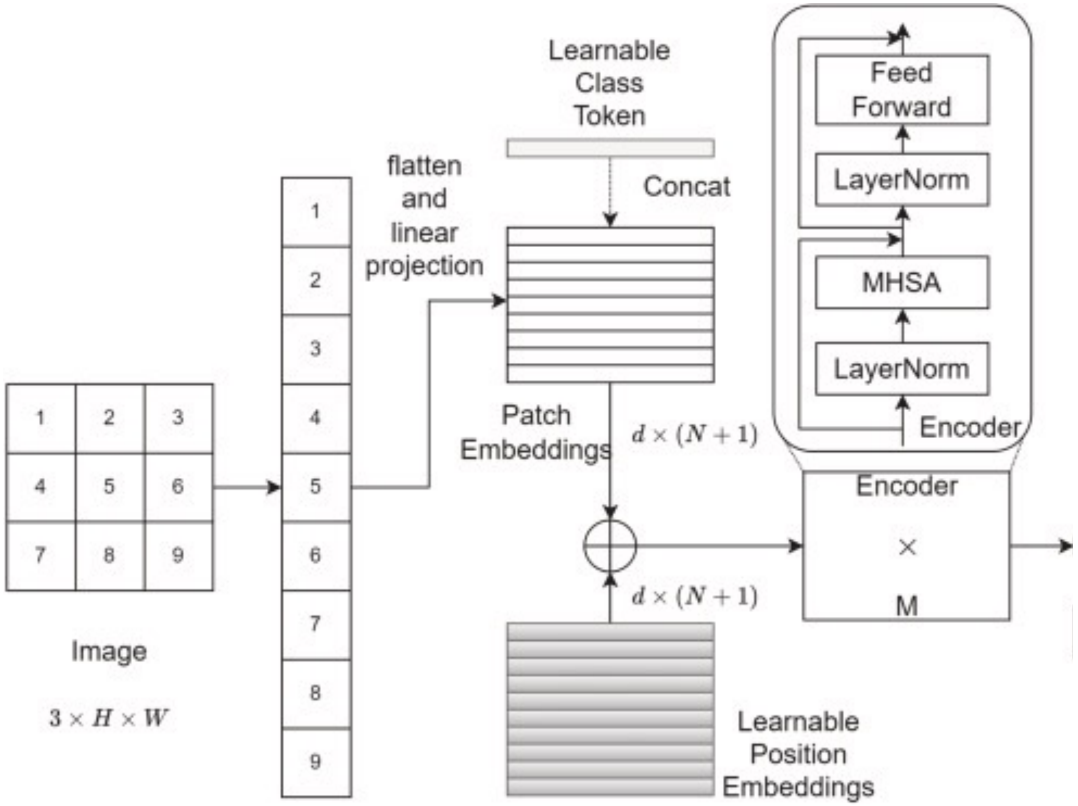
- [AGG] Token = [Class] Token → Z_{out}^{0}
  - Downstream task or Common space mapping
  - Learnable parameter
- GeLU Activation / Layer Normalization





3. Proposed Idea

3.4 The Transformer Architecture



(ViT) $\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\, \mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E};\, \cdots;\, \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$

(VATT) $\boldsymbol{z}_{\text{in}} = [\boldsymbol{x}_{\text{AGG}};\, \boldsymbol{x}_0\boldsymbol{W}_P;\, \boldsymbol{x}_1\boldsymbol{W}_P;\, \ldots;\, \boldsymbol{x}_N\boldsymbol{W}_P] + \boldsymbol{e}_{\text{POS}}$

3. Proposed Idea

3.5 Common Space Projection

- Common Space Projection → Contrastive learning
- FAC : MMV[5]에서 제안한 방법
    - Video - Audio : fine-grained space (512)
    - Video - Text : lower dimension → coarse-grained space (256)

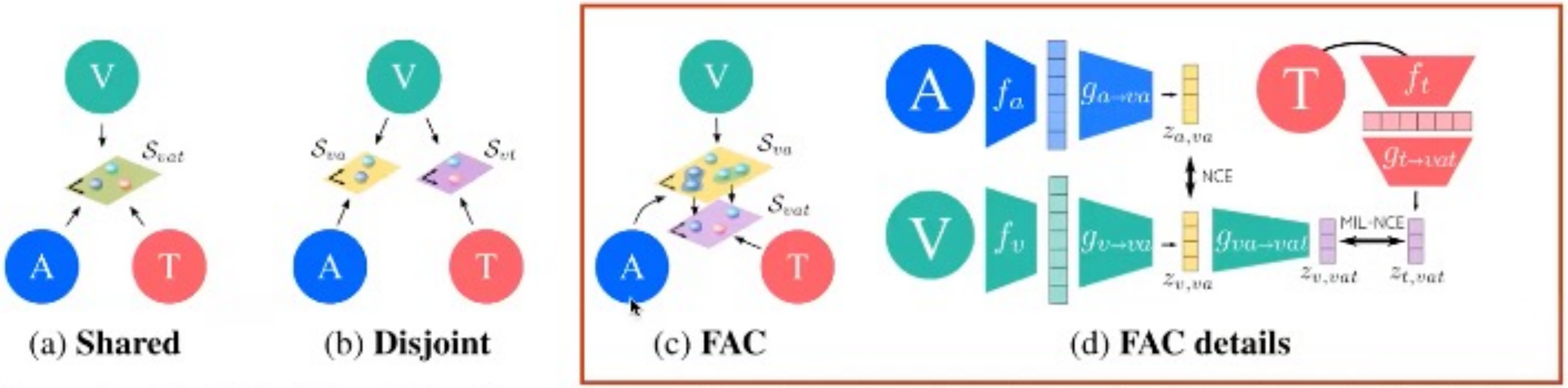
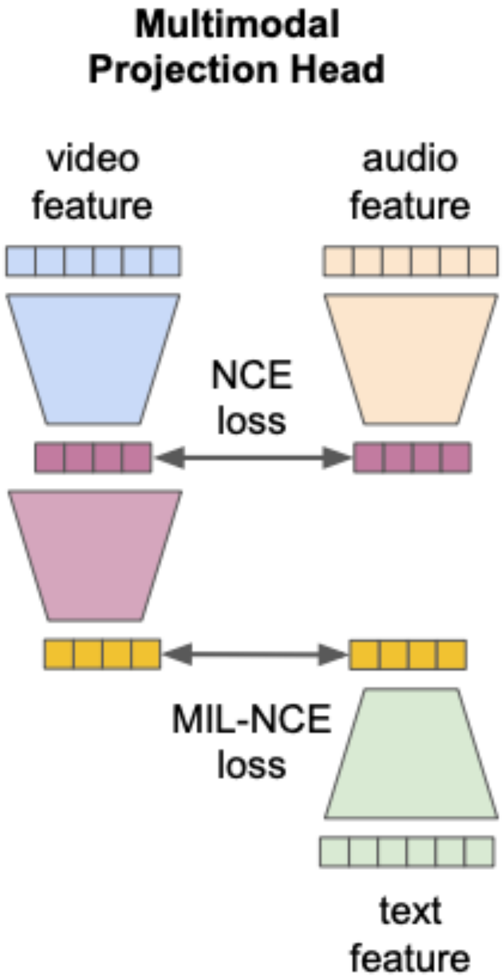


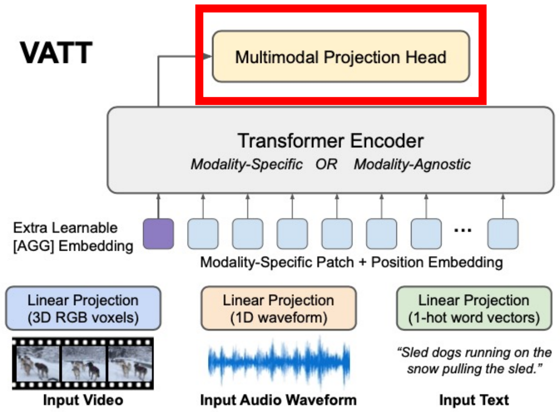
Figure 1: (a)-(c) Modality Embedding Graphs, (d) Projection heads and losses for the FAC graph. V=Vision, A=Audio, T=Text.

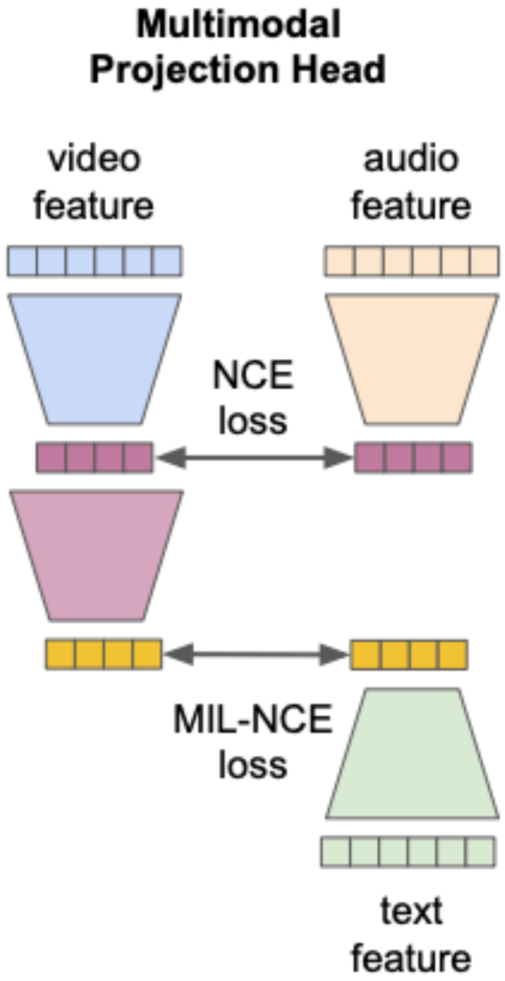


5) Alayrac, Jean-Baptiste, et al. "Self-supervised multimodal versatile networks." *Advances in Neural Information Processing Systems* 33 (2020): 25-37.

3. Proposed Idea

3.5 Common Space Projection





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

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

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

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

- 1 Layer Linear Projection
- ReLU + 2 Layer Linear Projection

- Batch Norm : After linear layer

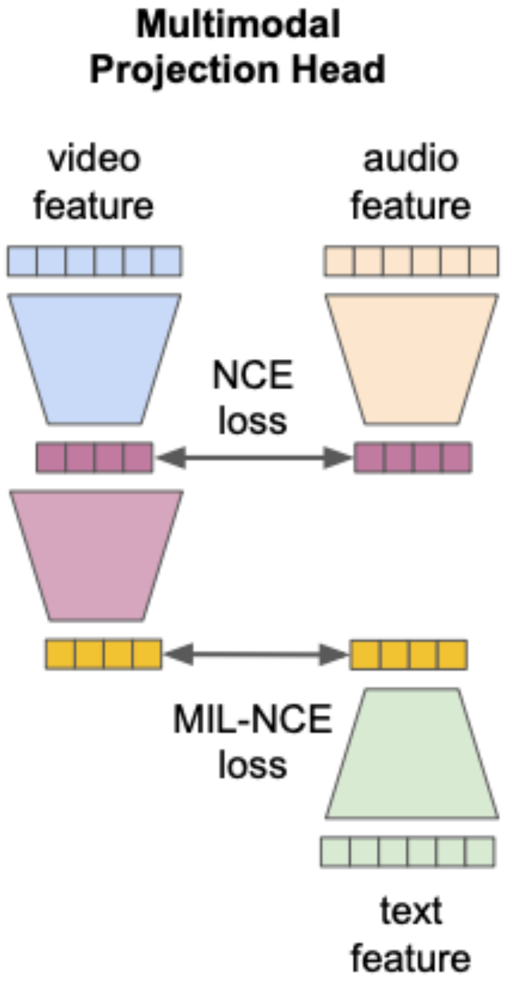
3. Proposed Idea

3.6 Multimodal Contrastive Learning

- Video - Audio : Noise-Contrastive Estimation (NCE loss)
- Video - Text : Multiple-Instance-Learning-NCE (MIL-NCE loss)

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

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



3. Proposed Idea

3.6 Multimodal Contrastive Learning

- (NCE loss)
  - Positive Pair (1개) → minimize
  - Negative pair → maximize
  - Cosine similarity [0,1]

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

3. Proposed Idea

3.6 Multimodal Contrastive Learning

- (MIL-NCE loss)
  - Positive Pair → minimize
  - Negative pair → maximize
  - Cosine similarity [0,1]

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

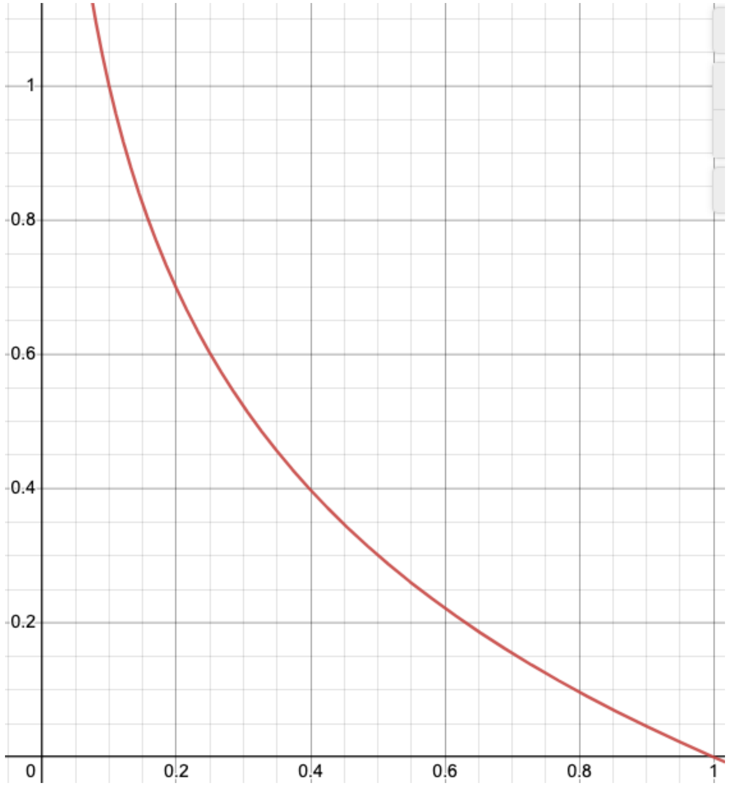
3. Proposed Idea

3.6 Multimodal Contrastive Learning

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

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

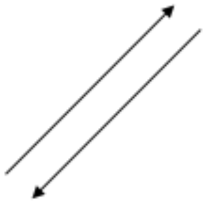
$$\mathcal{L} = \mathrm{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) + \lambda\text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}),$$



$\mathcal{N}$: number of negative pair

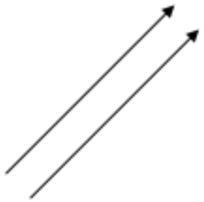
τ : temperature 변수

$\mathcal{P}$: 5 text clips



코사인 유사도 : -1

코사인 유사도 : 0



코사인 유사도 : 1

4. Evaluation

4.1 Setup

- Pre-train
  - AudioSet(Audio + Video + 0 Vector) + Howto100M(Video + Audio + Scripts)
    - Automatic Speech Recognition
  - Random Crop, Horizontal Flip, Color Augmentation, Normalize
  - Video
    - 32 x 224 x 224 x 3 (10fps)
    - Patch size : t=4 x h=16 x w=16 x 3
  - Audio
    - 48kHz
    - Patch size : 128
  - DropToken rate : 50%
  - Temperature 변수 τ : 0.07
  - λ : 1

4. Evaluation

4.1 Setup

- **Video action recognition :**
  - UFC101 (101 classes, 13,320 videos)
  - HMDB51 (51 classes, 6,766 videos)
  - Kinetics-400 (400 classes, 234,584 videos)
  - Kinetics-600 (600 classes, 366,016 videos)
  - Moments in Time (339 classes, 791,297 videos)

- **Audio event classification :**
  - ESC50 (50 classes, 2000 audio clips)
  - AudioSet (527 classes, 2M audio clips)

- **Zero-shot video retrieval :**
  - YouCook2 (3.1k video-text pairs)
  - MSR-VTT (1k video-text pairs)

- **Image classification :**
  - ImageNet-1000k (1000 classes, 1.2M)

Freeze & train linear classifier

4. Evaluation

4.1 Setup

- Network
  - Modality-agnostic
    - Medium Model (MA 155M parameter)
  - Modality-specific
    - (video-audio-text) backbone
    - Base-Base-Small (BBS 197M parameter)
    - Medium-Base-Small (MBS 264M parameter) : TPUv3 256개 3days
    - Large-Base-Small (LBS 415M parameter)

4. Evaluation

4.2 Results

- Fine-tuning for video action recognition
  - SOTA 달성
  - Agnostics = Base model
    - 단일 backbone 가능성
  - Model ↑
    - FLOPs ↑
      - $10^{12}/s$
    - Accuracy ↑

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.

4. Evaluation

4.2 Results

- Fine-tuning for audio event classification
  - SOTA 달성
  - Agnostics = Base model

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.

4. Evaluation

4.2 Results

- Fine-tuning for image classification
  - Train : Video
  - 다른 domain Transfer 가능성
    - Patch size : 4 x 16 x 16 x 3
    - Input image copy → 4장
  - Unlabeled data로 pretrain했지만 준수한 성능

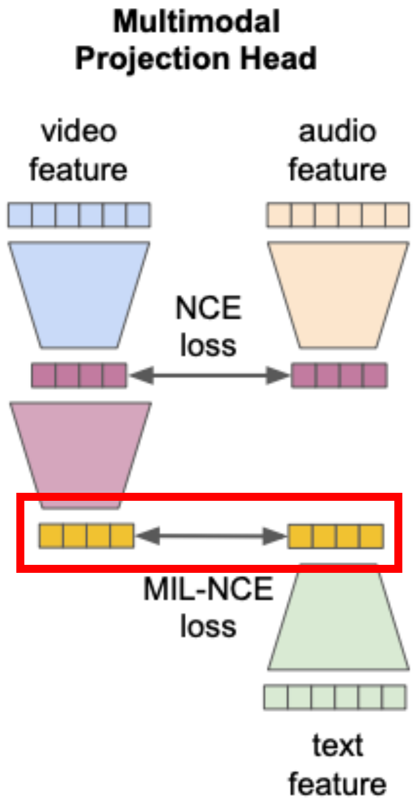
METHOD	PRE-TRAINING DATA	TOP-1	TOP-5
iGPT-L [16]	ImageNet	72.6	-
ViT-Base [25]	JFT	$\boxed{\textbf{79.9}}$	-
VATT-Base	-	64.7	83.9
VATT-Base	HowTo100M	$\boxed{78.7}$	93.9

Table 3: Finetuning results for ImageNet classification.

4. Evaluation

4.2 Results

- Zero-shot text-to-video retrieval
  - Zero-shot
    - Fine-tuning X
    - Semantic information
  - Metrics
    - Recall@10
    - MedR : True video 순위 median
  - S_{vt} space에서 representation 추출



$$Recall = \frac{TP}{TP + FN}$$

* 실제 True 값 중 model이 True라고 예측한 비율

4. Evaluation

4.2 Results

- Zero-shot text-to-video retrieval
  - SOTA X : noisy data
  - Batch size, epochs ↓ 비슷한 결과
    - Batch size : 8192
    - Epochs : 6
    - (YouCook2) MIL-NCE와 동일한 결과
    - (MSR-VTT) 성능 ↑
      - R@10 : 29.2
      - MedR : 42

			YouCook2		MSR-VTT	
METHOD	BATCH	EPOCH	R@10	MedR	R@10	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.

4. Evaluation

4.2 Results

- Feature Visualization
  - Modality-specific과 Modality-agnostic 비교
    - Fine-tune : better separation
    - Specific과 agnostic과는 명확한 차이 X

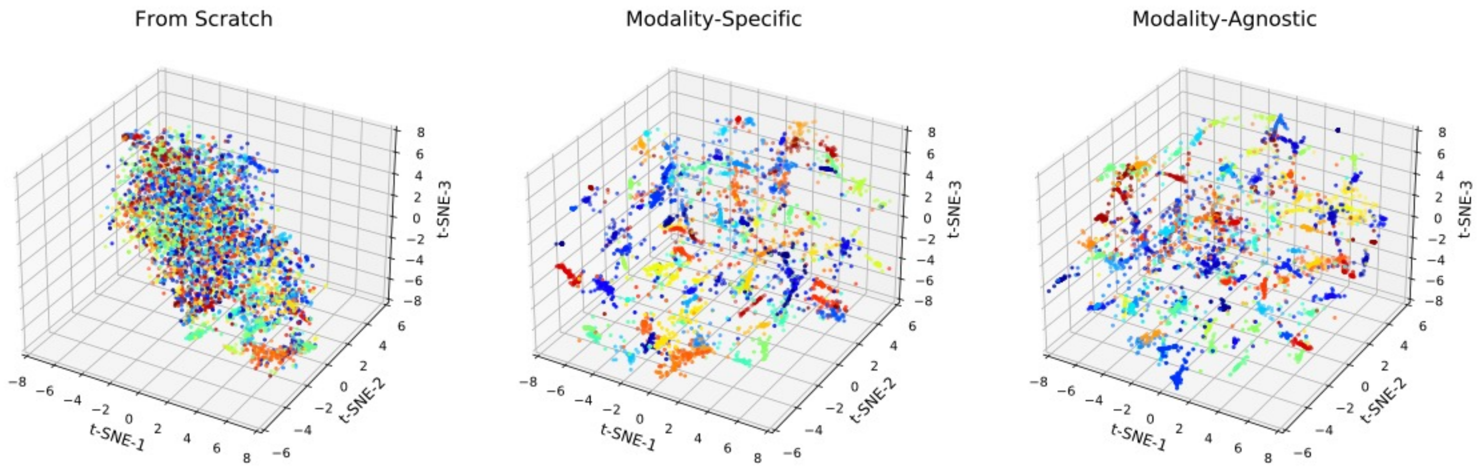
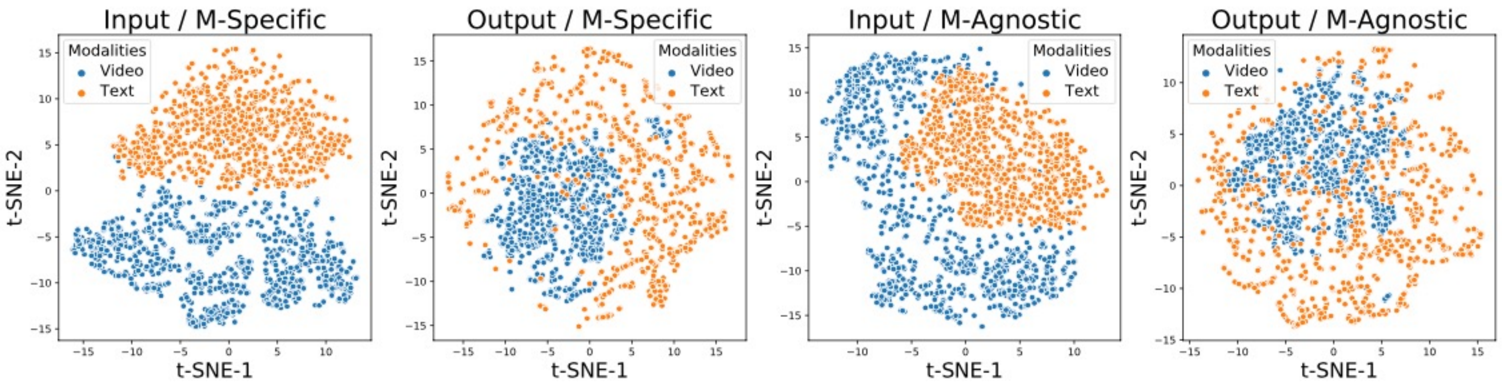


Figure 2: t-SNE visualization of the feature representations extracted by the vision Transformer in different training settings. For better visualization, we show 100 random classes from Kinetics-400.

4. Evaluation

4.2 Results

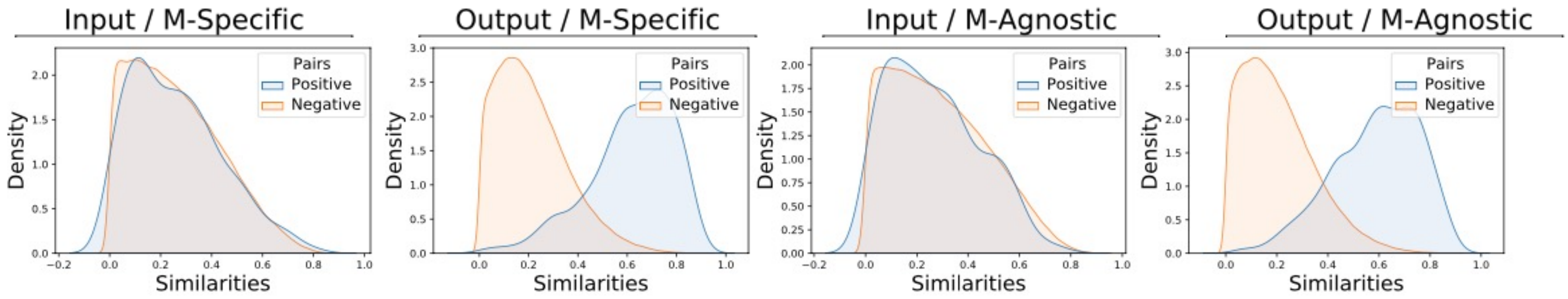
- Feature Visualization
  - After Tokenization layer와 After Common space 비교
  - Agnostic이 좀 더 섞여 있는 모습
    - 같은 concept을 묘사하는 다른 symbol을 다른 modality로 간주 → 여러 언어를 지원하는 NLP 모델과 유사



4. Evaluation

4.2 Results

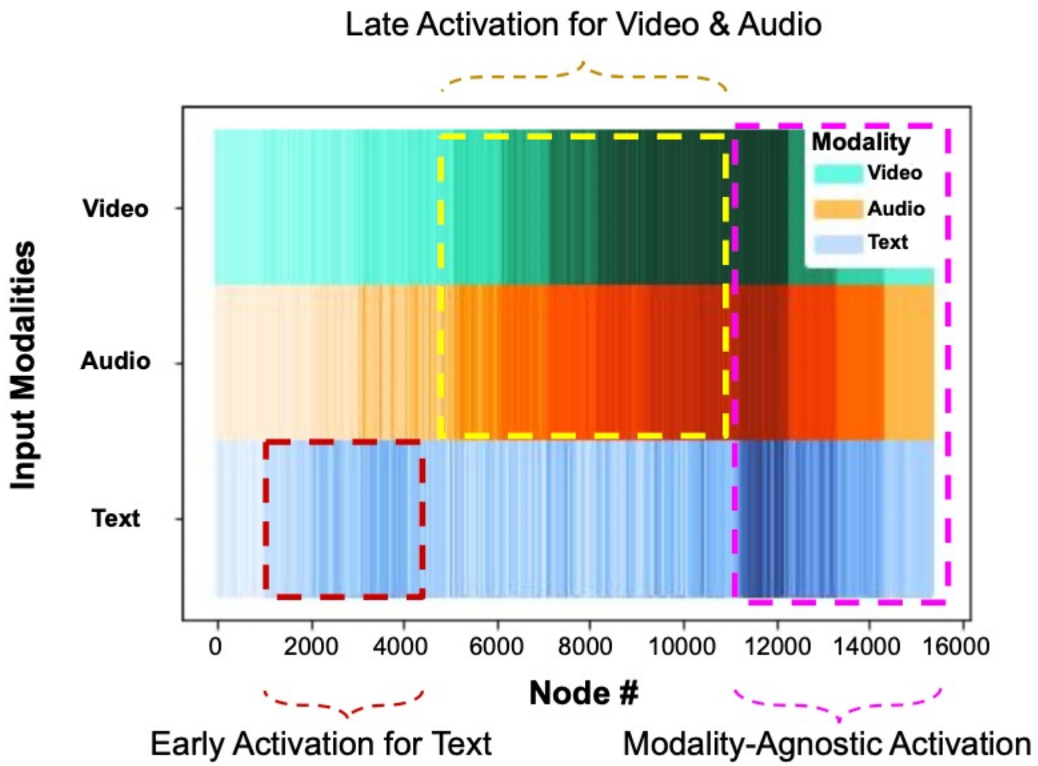
- Feature Visualization
  - After Tokenization layer와 After Common space 비교
  - Positive pair와 negative pair 분포는 비슷



4. Evaluation

4.2 Results

- Model Activations
  - VATT average activation of the modality-agnostic
    - Text : early node activated
    - Video and audio : middle to later node activat
    - All modality : last layer activated
  - Mixture of Experts 가능성



4. Evaluation

4.2 Results

- Effect of Drop Token
  - DropToken이 downstream과 pre-train에 미치는 영향
    - Pre-train DropToken rate : 75%, 50%, 25%, 0%
    - Accuracy와 Cost의 절충안 : 50%

	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

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

4. Evaluation

4.2 Results

- Effect of Drop Token
  - 50% Pre-train model → fine-tune
    - DropToken rate : 75%, 50%, 25%, 0%
  - DropToken Vs low-resolution

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

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

5. Conclusion & Future Work

5.1 Conclusion

- Transformer 기반 Self-supervised Multi-modal Representation Learning Framework
    - Weight을 share해도 Representation을 학습 가능 (Agnostic)
    - Labeled data 의존도↓

5.2 Future Work

- Modality-agnostic model
- Computational Cost↑

감사합니다.

육현준