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

School of Industrial and Management Engineering, Korea University

Jae Hoon Kim

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text (arXiv, 2021.4)

- Microsoft에서 연구하였으며 2021년 11월 17일 기준으로 19회 인용됨

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

Hassan Akbari* [1,2], Liangzhe Yuan[1], Rui Qian* [1,3], Wei-Hong Chuang[1], Shih-Fu Chang[2], Yin Cui[1], Boqing Gong[1]
[1]Google [2]Columbia University [3]Cornell University
{lzyuan,whchuang,yincui,bgong}@google.com {ha2436,sc250}@columbia.edu {rq49}@cornell.edu

Abstract

We present a framework for learning multimodal representations from unlabeled data using convolution-free Transformer architectures. Specifically, our Video-Audio-Text Transformer (VATT) takes raw signals as inputs and extracts multimodal representations that are rich enough to benefit a variety of downstream tasks. We train VATT end-to-end from scratch using multimodal contrastive losses and evaluate its performance on video action recognition, audio event classification, image classification, and text-to-video retrieval. Furthermore, we study a modality-agnostic, single-backbone Transformer by sharing weights among the three modalities. We show that the convolution-free VATT outperforms state-of-the-art ConvNet-based architectures in the downstream tasks. Especially, VATT's vision Transformer achieves the top-1 accuracy of 82.1% on Kinetics-400, 83.6% on Kinetics-600, and 41.1% on Moments in Time, new records while avoiding supervised pre-training. Transferring to image classification leads to 78.7% top-1 accuracy on ImageNet compared to 64.7% by training the same Transformer from scratch, showing the generalizability of our model despite the domain gap between videos and images. VATT's audio Transformer also sets a new record on waveform-based audio event recognition by achieving the mAP of 39.4% on AudioSet without any supervised pre-training. VATT's source code is publicly available.[1]

ral networks [47, 8] and CNNs [110, 36], to more general architectures constructed upon self-attention. Particularly, Transformers [93] become the de facto model architecture for NLP tasks [27, 76, 77, 11]. Pre-training a Transformer on large text corpora followed by fine-tuning gives rise to state-of-the-art results for different downstream tasks.

In view of the success of the attention mechanism in NLP, there has been a rich line of works exploring its potential in computer vision. Early work studied hybrid models consisting of both convolutions and attention modules [94, 100, 40, 111]. Recent studies showed that convolution-free, specially designed all-attention models can match CNNs' performance on image recognition tasks [112, 48, 79]. Most recently, Dosovitskiy *et al.* [29] achieved impressive performance on several image recognition tasks, including ImageNet [26], using a pre-trained Transformer with minimal architecture changes. Their work delivered a compelling message that "large scale (supervised) training trumps inductive bias (for image classification)." This conclusion was further extended to video recognition tasks by [10, 6].

However, the large-scale supervised training of Transformers is essentially troubling for two main reasons. First, it rules out the much larger other part of "big visual data," *i.e.*, the vast amount of unlabeled, unstructured visual data. As a result, the supervised training strategy could produce biased systems that require even more labeled data to correct their biases. Second, this strategy fundamentally limits

Research Purpose

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

- Multi-modal 데이터에 대하여 self-supervised learning 방법론을 적용
- 최근 텍스트, 시계열 뿐만 아니라 이미지에도 활용되고 있는 Transformer를 활용하여 modal-agnostic한 multi-modal모델 제안 (단, 주요 실험은 modality-specific으로 진행함)

Multi-modality

Video

Audio

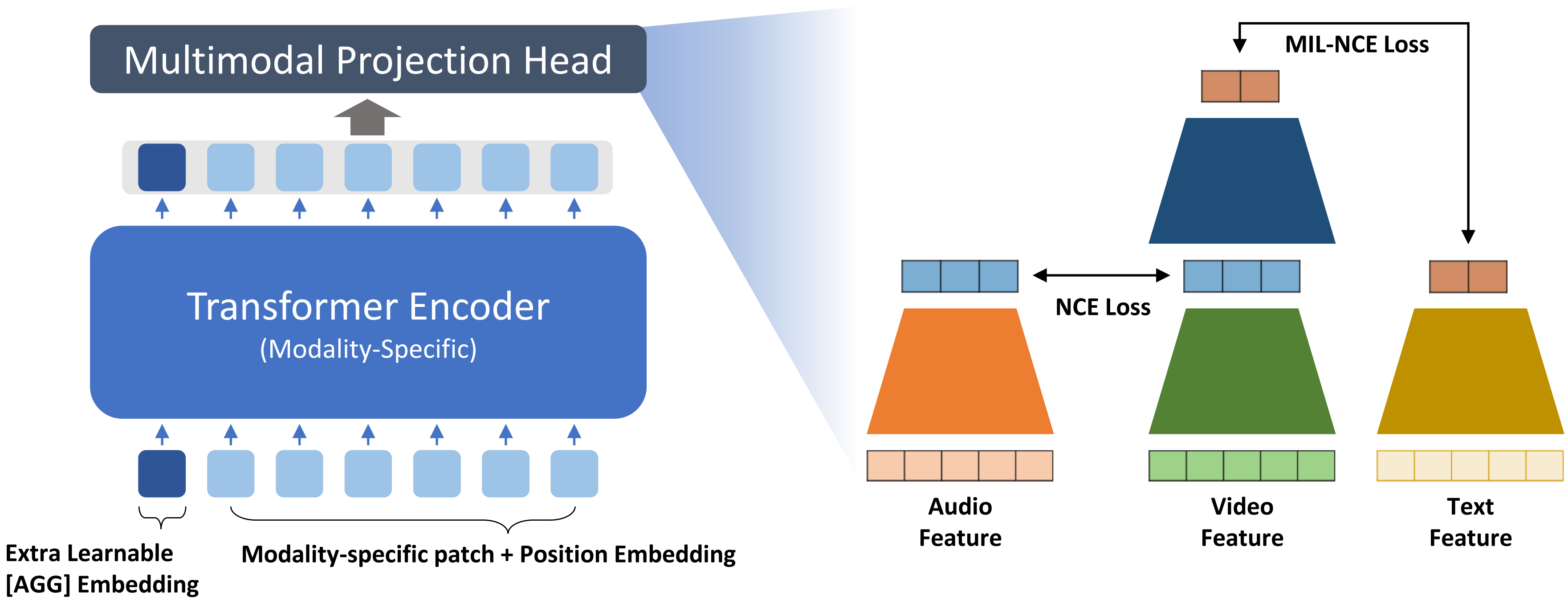
Text

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

Video, Audio, Text Transformers (VATT)

❖ Diagram of VATT (overall architectures)

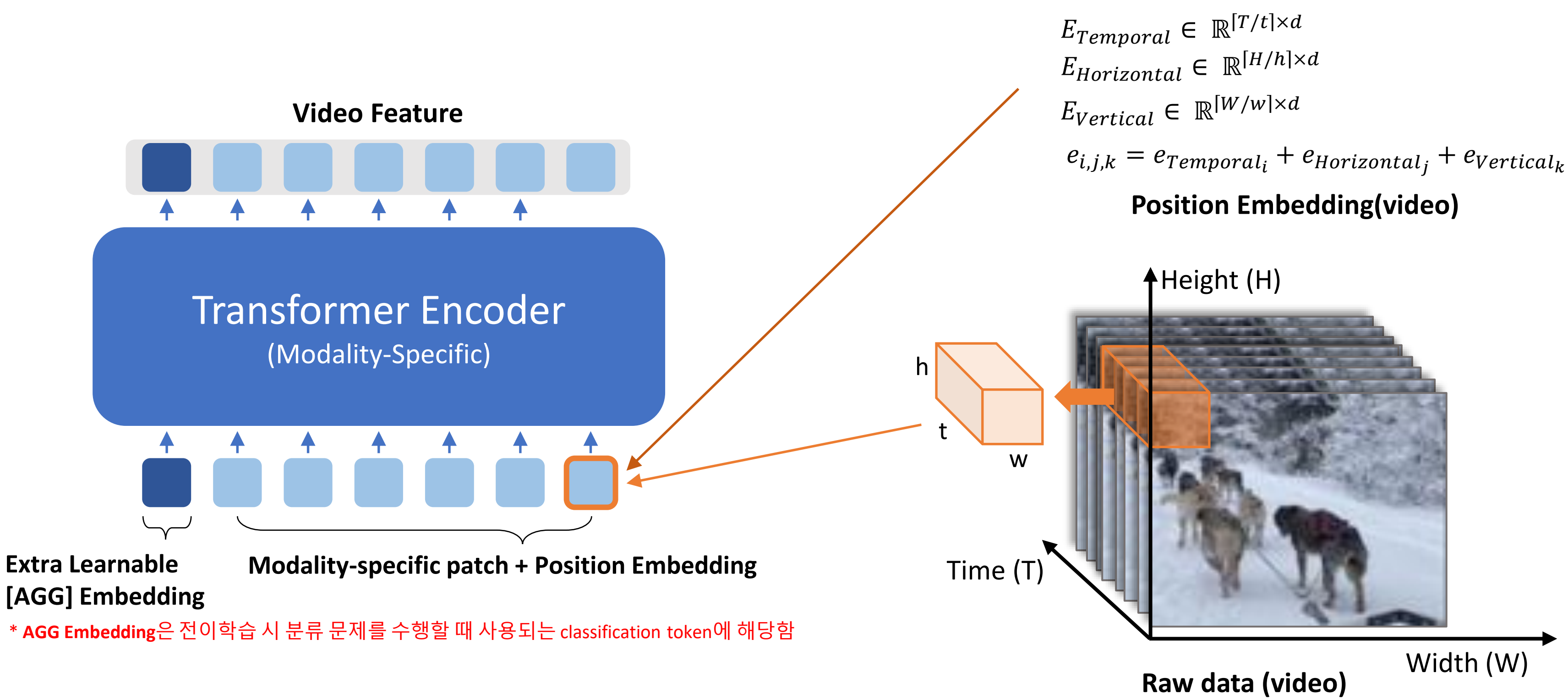
- VATT는 modality-agnostic 혹은 modality-specific 방식으로 구현될 수 있음
- 논문에서는 modality-specific 로 진행하였으며 modal 별로 총 세 개의 Transformer를 생성함
- Transformers 인코더는 A. Vaswani(2017)[1]가 제안한 구조를 사용함



1. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

Video, Audio, Text Transformers (VATT)

- Diagram of VATT (Video Encoder)
  - 각 패치는 voxel 단위로서 time * (height * weight * 3 channels)으로 구성됨
    - ✓ 각 패치는 linear projection을 통해 d 차원의 벡터로 표현됨
  - 각 시간대 및 영상 패치 위치에 대한 positional encoding을 수행함



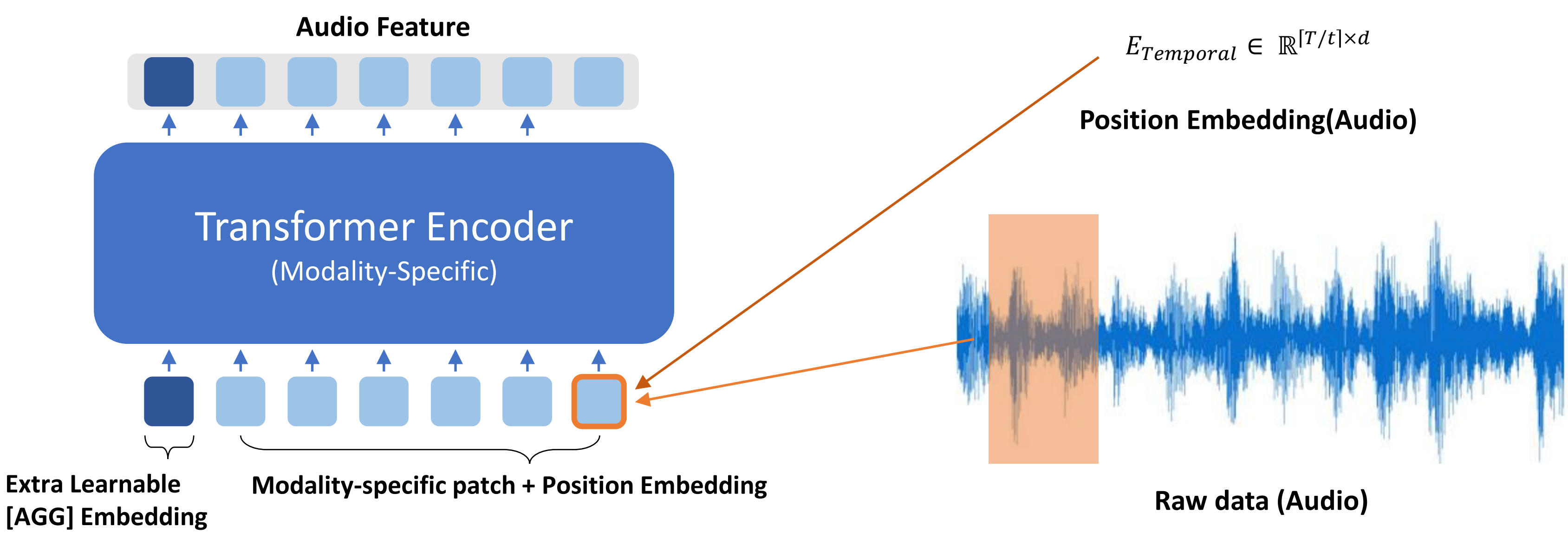
* **AGG Embedding**은 전이학습 시 분류 문제를 수행할 때 사용되는 classification token에 해당함

DMQA

Video, Audio, Text Transformers (VATT)

❖ Diagram of VATT (Audio Encoder)

- 각 패치는 t 만큼의 길이로 나눈 audio waveform에 해당함
  - ✓ 각 패치는 linear projection을 통해 d 차원의 벡터로 표현됨
- 각 나눈 구간의 위치에 대한 positional encoding을 수행함

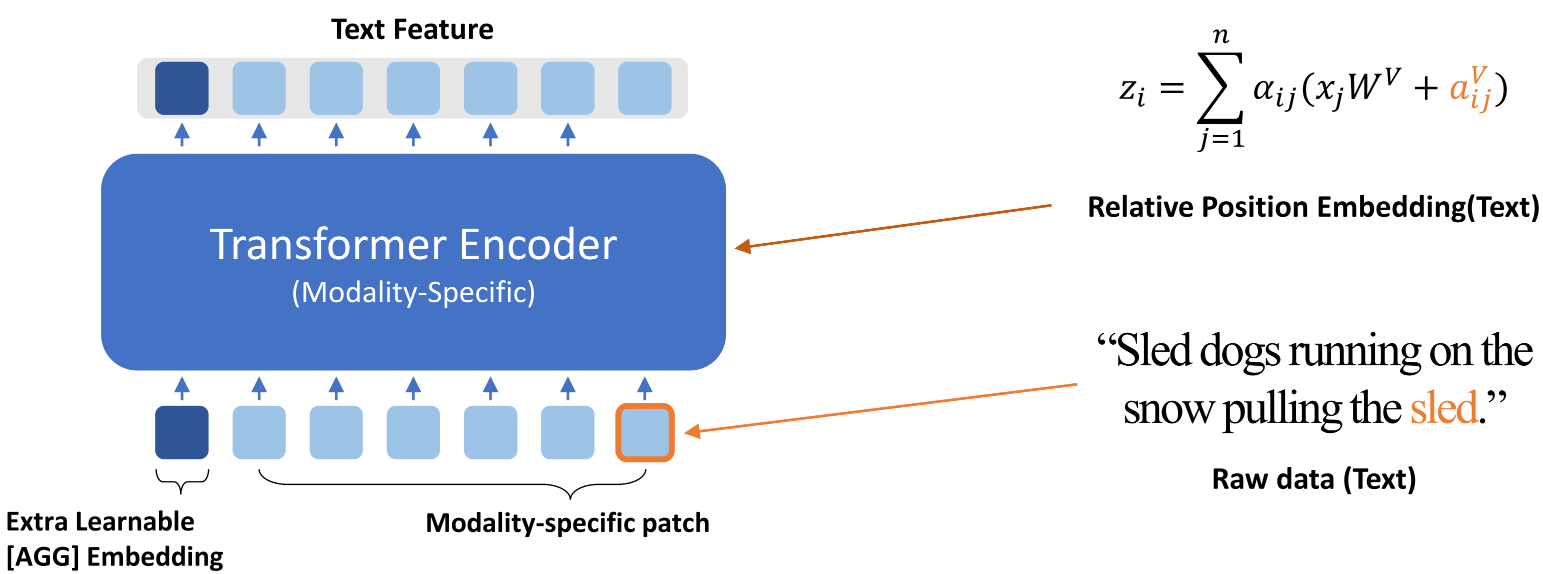


* AGG Embedding은 전이학습 시 분류 문제를 수행할 때 사용되는 classification token에 해당함

Video, Audio, Text Transformers (VATT)

❖ Diagram of VATT (Text Encoder)

- 학습 데이터셋에 등장하는 단어에 대해서 embedding table을 생성함
- Text 데이터에 대해서는 relative positional encoding을 수행함
  - ✓ 기존의 positional encoding은 입력 데이터에 적용되며 절대적인 위치를 임베딩
  - ✓ Relative positional encoding은 attention 연산과 함께 수행되며 연산 되는 토큰의 위치를 중심으로 다른 토큰 간의 상대적인 위치를 임베딩



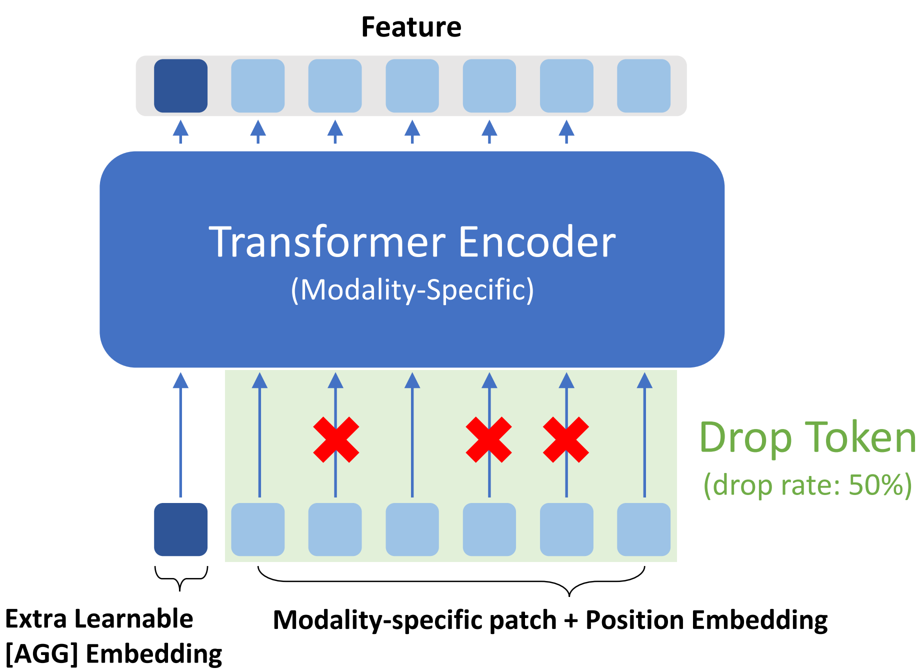
$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + a_{ij}^V)$$

* AGG Embedding은 전이학습 시 분류 문제를 수행할 때 사용되는 classification token에 해당함

Video, Audio, Text Transformers (VATT)

❖ Diagram of VATT (Drop Token)

- Video와 Audio modality의 경우 입력 데이터의 일부를 탈락시키는 기법을 사용함
- Transformer의 학습 시 계산 복잡도가 입력 데이터 길이의 제곱에 비례하기 때문
  - ✓ Video와 audio는 text에 비해서 특징이 더 많기 때문에 데이터의 길이가 더 길어질 수밖에 없음
- Drop Token을 적용할 경우 낮아지는 계산 복잡도에 비해 성능이 거의 유지되는 모습을 보임



	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

Table 13. Linear classification top-1 accuracy vs. sampling rate vs. inference GFLOPs in the Medium-Base-Small (MBS) setting.

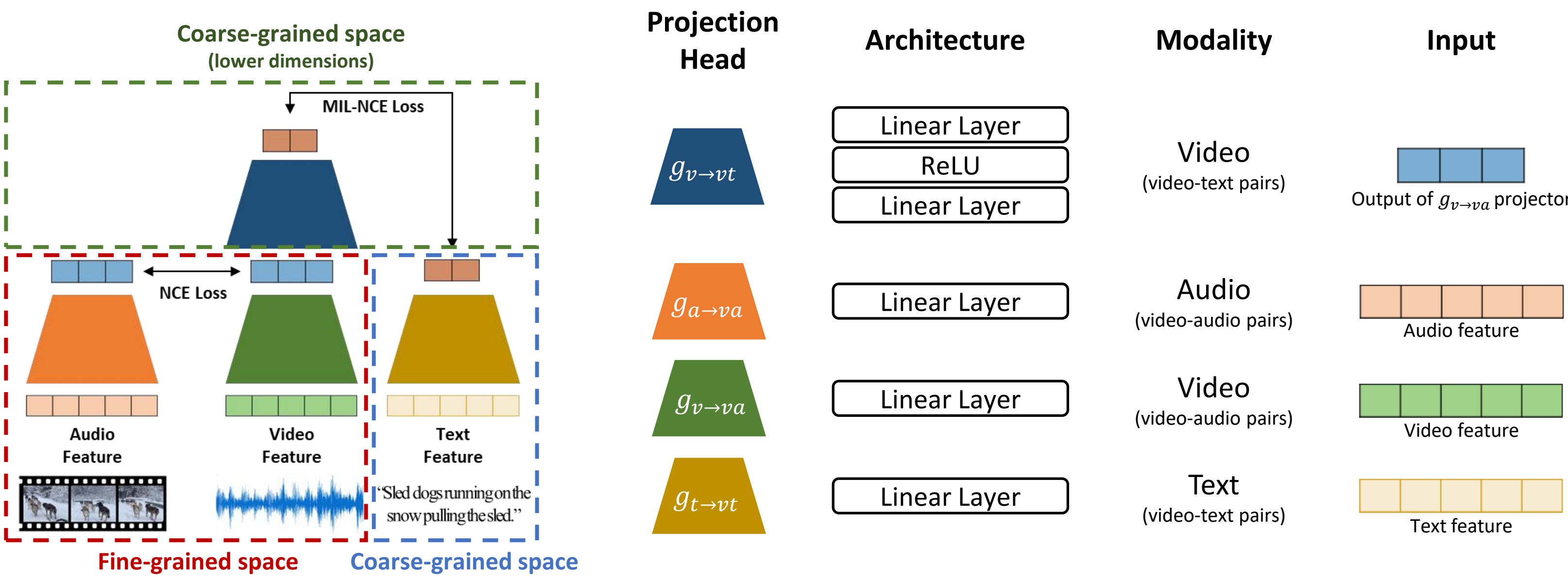
**** Drop Rate 0% → 50% 시 (논문 실험 세팅 값)**

- ✓ 계산 복잡도 약 52.1% 하락
- ✓ 모델 성능 평균 약 1.9% 하락

Video, Audio, Text Transformers (VATT)

❖ Diagram of VATT (Common Space Projection)

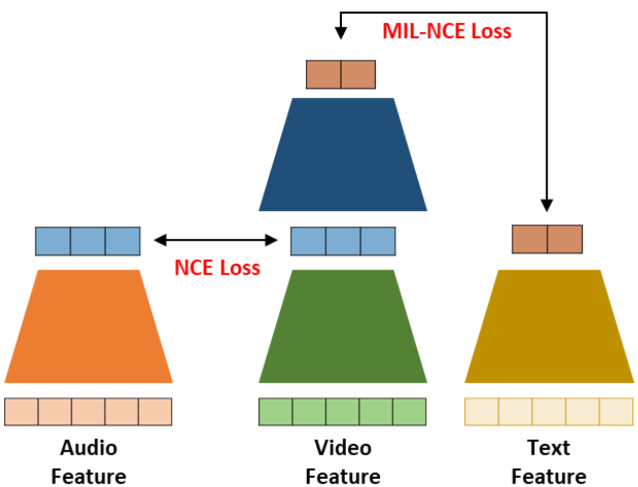
- Video, Audio의 데이터 밀도와 Text 데이터 밀도에는 차이가 존재함
  - ✓ Video & Audio: 하나의 토큰에 여러 time step의 특징이 들어감 → Fine-grained feature space
  - ✓ Text: 하나의 토큰에 하나의 단어(특징)가 들어감 → Coarse-grained feature space
- 따라서 loss를 계산할 modality 간의 feature space 수준을 맞춰줄 필요가 있음



Video, Audio, Text Transformers (VATT)

❖ Diagram of VATT (Multi-modal Contrastive Learning)

- Video-audio pairs에는 Noise-Contrastive Estimation (NCE) Loss를 적용함
- Video-text pairs에는 Multiple-Instance-Learning-NCE (MIL-NCE) Loss를 적용함
  - ✓ 하나의 video feature에 대해서 다수의 text features에 대한 유사도를 계산함
  - ✓ Video의 한 장면이 한 문장과 시간 순서 측면에서 완벽하게 일치하지 않기 때문에 해당 시간대의 주변 문장을 모두 활용



Total Loss

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

* λ는 loss의 비중을 조절함

NCE Loss

$$NCE(z_{v,va}, z_{a,va}) = -\log\left(\frac{\exp\left(\frac{z_{v,va}^T z_{a,va}}{\tau}\right)}{\sum_{i=1}^{B} \exp\left(\frac{{z_{v,va}^i}^T z_{a,va}^i}{\tau}\right)}\right)$$

MIL-NCE Loss

$$MIL - NCE(z_{v,vt}, \{z_{t,vt}\}) = -\log\left(\frac{\sum_{z_{t,vt} \in P(z_{v,vt})} \exp\left(\frac{z_{v,vt}^T z_{t,vt}}{\tau}\right)}{\sum_{z_{t,vt} \in P(z_{v,vt}) \cup \mathcal{N}(z_{v,vt})} \exp\left(\frac{{z_{v,vt}^i}^T z_{t,vt}^i}{\tau}\right)}\right)$$

Experiments

❖ Model & Data settings

- 사전학습은 온라인에 업로드 된 비디오로 수행되었으며 모두 레이블이 달려있지 않음
- 전이학습 및 검증은 각 modality에서 사용되는 벤치마크 데이터셋이 사용되었음

Default pre-training settings (all models)

- Dataset: Online video dataset with audio and text (136M)
- Optimizer: Adam
- Batch size: 2048
- Initial learning rate: 0.0001
- Warm-up steps: 10,000
- Cosine learning scheduler

Default architecture settings (all models)

Model	Layers	Hidden Size	MLP Size	Heads	Params
Small	6	512	2048	8	20.9 M
Base	12	768	3072	12	87.9 M
Medium	12	1024	4096	16	155.0 M
Large	24	1024	4096	16	306.1 M

Table 1. Details of the Transformer architectures in VATT.

DMQA

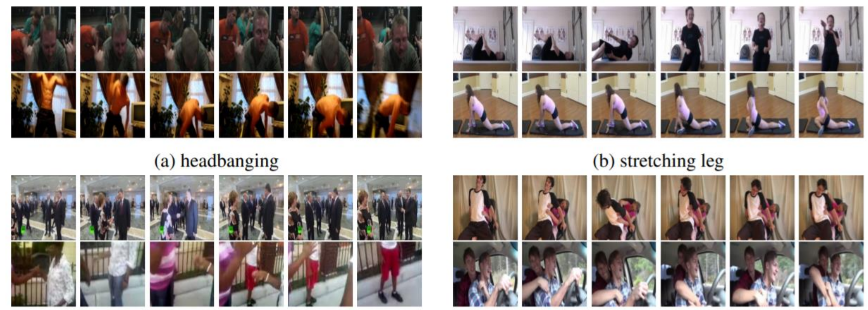
Experiments

❖ Fine-tuning for video action recognition

- Modality-agnostic 방식은 하나의 Transformer backbone을 가지는 모델
- 기존의 모든 video action recognition 방법론에 비해서 좋은 성능을 모임

Kinetics-400 Dataset

* 사람 행동을 기록한 영상 데이터로 총 400가지의 클래스를 가지고 있음



(a) headbanging (b) stretching leg (c) shaking hands (d) tickling

Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

METHOD	TOP-1	TOP-5	TFLOPS
ARTNet [98]	69.2	88.3	6.0
I3D [16]	71.1	89.3	-
R(2+1)D [30]	72.0	90.0	17.5
MFNet [60]	72.8	90.4	-
Inception-ResNet [2]	73.0	90.9	-
bLVNet [32]	73.5	91.2	0.84
A^2-Net [22]	74.6	91.5	-
TSM [61]	74.7	-	-
S3D-G [102]	74.7	93.4	-
Oct-I3D+NL [21]	75.7	-	0.84
D3D [88]	75.9	-	-
GloRe [23]	76.1	-	-
I3D+NL [98]	77.7	93.3	10.8
ip-CSN-152 [92]	77.8	92.8	-
MoViNet-A5 [51]	78.2	-	0.29
CorrNet [17]	79.2	-	6.7
LGD-3D-101 [75]	79.4	94.4	-
SlowFast [34]	79.8	93.9	7.0
X3D-XXL [33]	80.4	94.6	5.8
TimeSFormer-L [10]	80.7	94.7	7.14
VATT-Base	79.6	94.9	9.09
VATT-Medium	81.1	**95.6**	15.02
VATT-Large	**82.1**	95.5	29.80
VATT-MA-Medium	79.9	94.9	15.02

Supervised 방식으로 사전학습된 ViT 모델에 전이학습을 수행한 모델
→ **VATT는 레이블을 사용하지 않는 사전학습 방식으로써 더 나은 성능을 보임**

Modality-specific

Modality-agnostic

하나의 Transformer backbon만으로도
세 개의 backbon을 사용한 경우와 유사한 성능을 보임
→ **Modality를 통합하여 학습한 backbone도 충분히 사용 가능함을 시사**

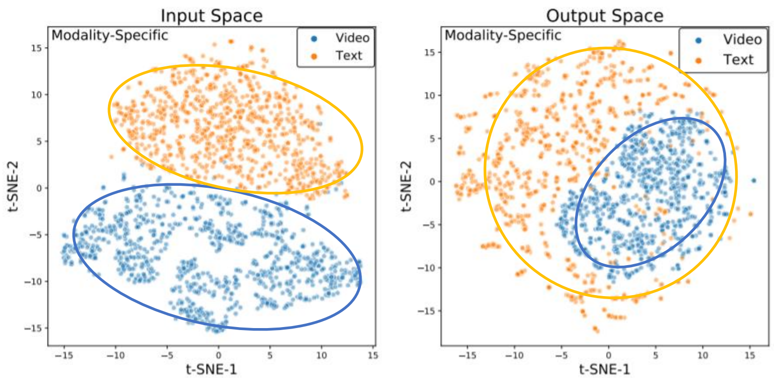
DMQA

Experiments

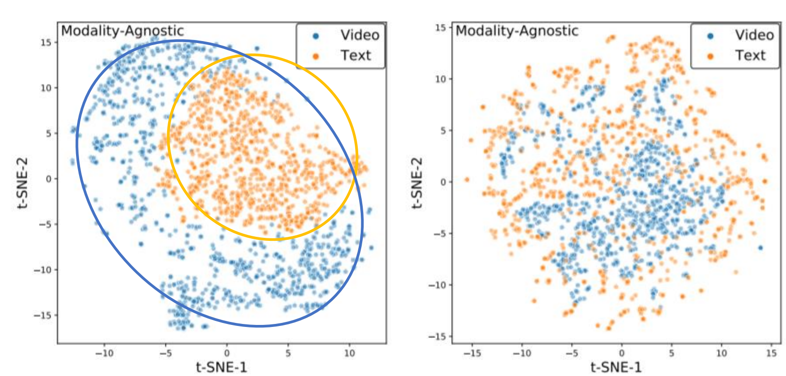
❖ Feature visualization

- Modality-agnostic 모델과 modality-specific 모델에서 추출한 데이터의 특징을 비교함
- 입력데이터의 경우 두 모델 모두 video와 text의 특징이 서로 다른 공간에 군집하고 있음
- 반면 모델로부터 추출한 특징의 경우 두 modality가 서로 다른 양상을 보여줌
  - ✓ Modality-**agnostic**: 두 modality가 특징 공간에서 **혼재**되어 있는 모습을 보임
  - ✓ Modality-**specific**: 두 modality가 특징 공간에서 **분리**되어 있는 모습을 보임

➡ 이는 **modality-agnostic** 모델은 **서로 다른 modality를 가진 데이터라도 그 의미(semantic)를 같게 표현**하는 것으로 해석할 수 있음



Modality-Specific



Modality-Agnostic

https://littlefoxdiary.tistory.com/85

Conclusion

❖ Conclusion & Limitations

- Transformer로 수행할 수 있는 self-supervised multimodal learning 프레임워크를 제안함

- Inductive bias가 약하기 때문에 하나의 네트워크가 다양한 modality를 잘 학습할 수 있으며 modality-specific 모델과의 성능 차이가 크지 않음을 보여줌

- Drop Token 기법으로 일정 수준 해결하였지만 근본적으로 Vanilla Transformer는 많은 연산량과 메모리를 요구하는 단점이 있음

DMQA

Reference

1. Akbari, Hassan, et al. "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text." *arXiv preprint arXiv:2104.11178* (2021).
2. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
3. Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).
4. https://littlefoxdiary.tistory.com/85
5. https://robot-vision-develop-story.tistory.com/29

Thank You