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

Hassan Akbari, Wei-Hong Chuang, Liangzhe Yuan, Shih-Fu Chang, Boqing Gong, Rui Qian, and Yin Cui

NeurIPS 2021

Presented by Luchao Qi & Myles Mason

Motivation

Goal:

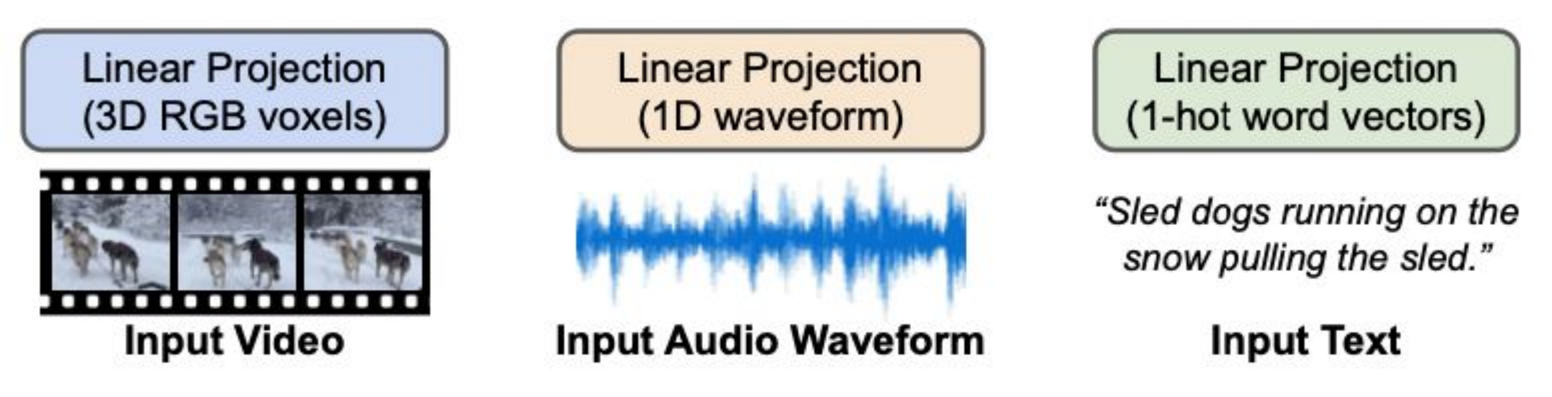
Develop a structure for learning multimodal representations from **unlabeled data** with a convolution free transformer architecture **from scratch**

Why raw signals?

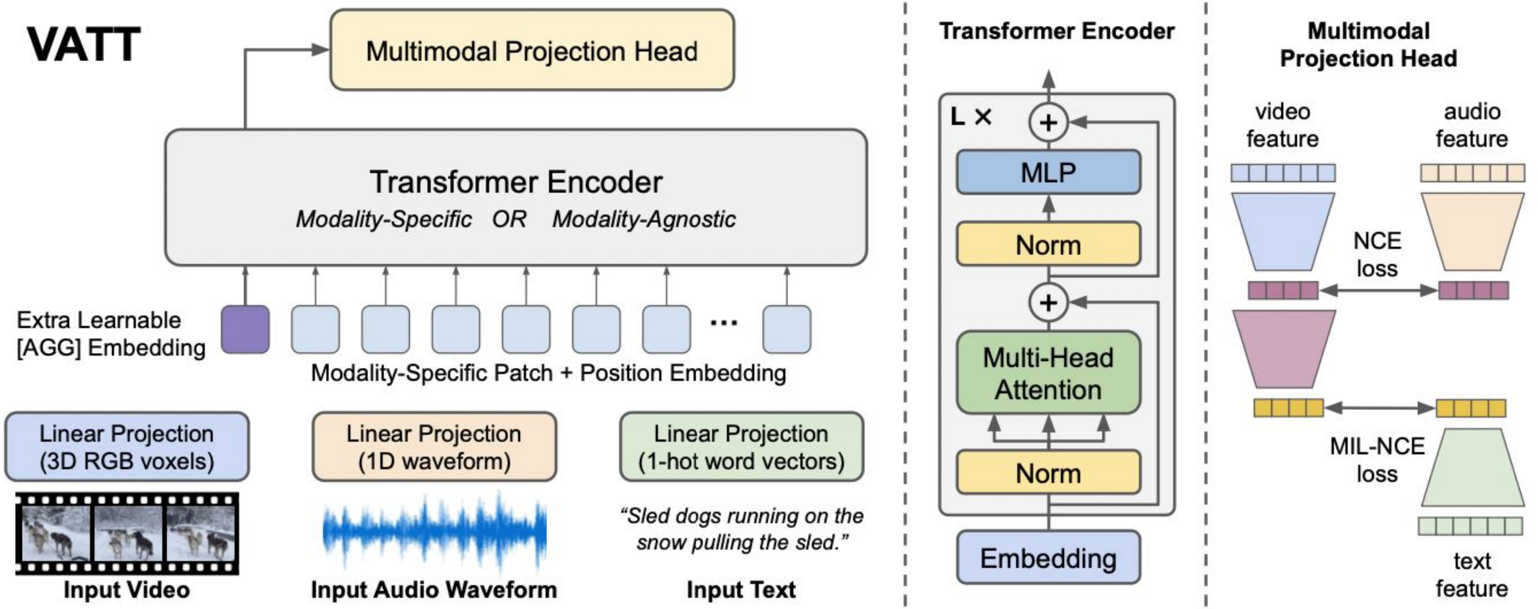
- Transformers are labeled-data hungry
- High costs for labeled data acquisition
    - Remember the paper battle back to Monday - Large noisy data vs. clean small data?

Can we extract all information from a video clip?

- Given an input video, audio waveform, or text we want to extract high level feature information as the aggregated representation of the whole input.



Yes we can, introducing VATT



Modality-agnostic - the idea is to test whether there exists a single, general-purpose model for all the modalities

Tokenization and Positional Encoding: Video

Learnable Weight

$$\boldsymbol{W}_{vp} \in \mathbb{R}^{t \cdot h \cdot w \cdot 3 \times d} \quad \boldsymbol{W}_{ap} \in \mathbb{R}^{t' \times d} \quad \boldsymbol{W}_{tp} \in \mathbb{R}^{v \times d}$$

Voxel 3D Positional Embedding Vector

Voxel dimensional vectors

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

$$\boldsymbol{E}_{\text{Temporal}} \in \mathbb{R}^{\lceil T/t \rceil \times d}, \quad \boldsymbol{E}_{\text{Horizontal}} \in \mathbb{R}^{\lceil H/h \rceil \times d}, \quad \boldsymbol{E}_{\text{Vertical}} \in \mathbb{R}^{\lceil W/w \rceil \times d}$$

Temporal Positional Embedding

Horizontal Positional Embedding

Vertical Positional Embedding

DropToken

Redundancies information in different modalities (audio/video)

Since the Transformer's computational complexity is quadratic $O(N^2)$ where N is the number of tokens in the input sequence.

- Sample a portion of the tokens and then feed the sampled sequence, not the complete set of tokens, to the transformer

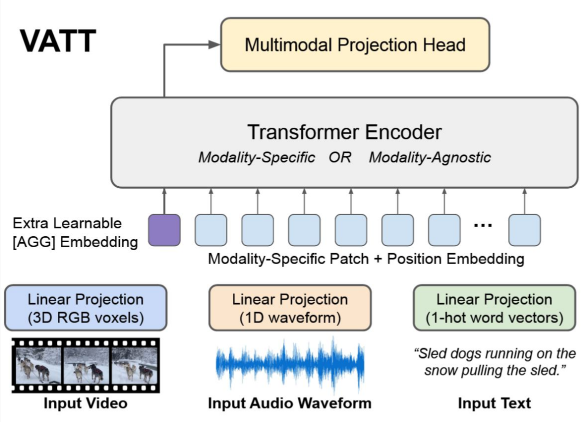
Common space projection - misalignment for noisy multi-modality data

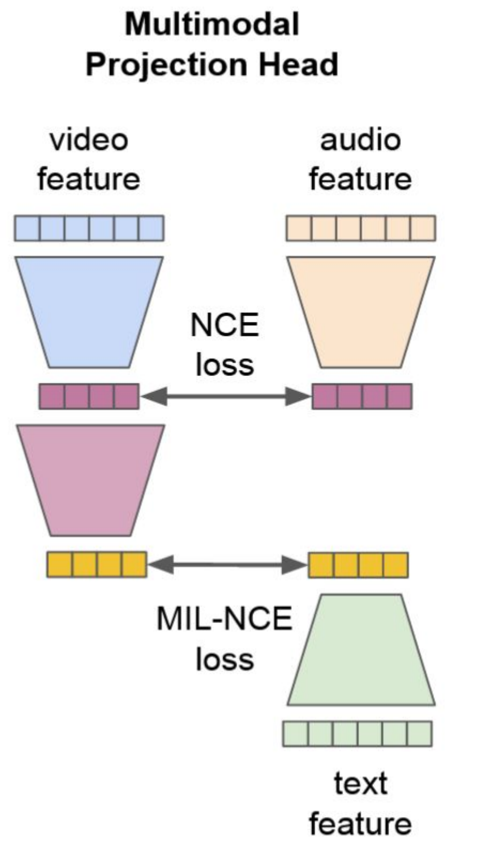
Cross-modality regularization

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

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

- such comparison is more feasible if we assume there are different levels of semantic granularity for different modalities





MIL-NCE loss

Multiple Instance Learning Noise Contrastive Estimation

- First proposed from paper presented on Monday: End-to-End Learning of Visual Representations

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

Video-audio pairs

Video-text pairs

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

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

Downstreamed Task

- Downstream on four tasks
  - Video action recognition
  - Audio event classsificaition
  - Text-to video retrieval
  - Image classification
- Pretraining on AudioSet and **HowTo100M**
  - video-audio pairs from AudioSet
  - video-audio-text triplets from HowTo100M
- Finetuning on other datasets OR zero-shot depending on the downstreaming task

Fine-tuning for video action recognition

Fine-tune VATT's vision Transformer on Kinetics-400, Kinetics-600, and Moments in Time

modality-agnostic backbone (VATT-MA-Medium)

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 7: Details of the Transformer architectures in VATT.

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

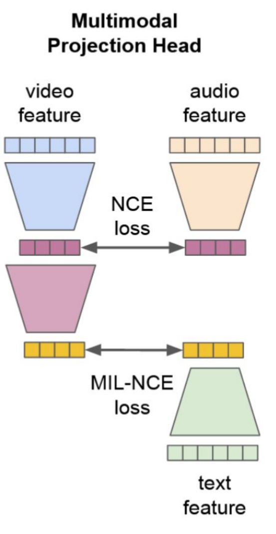
Table 1: Video action recognition accuracy on Kinetics-400, Kinetics-600, and Moments in Time.

[9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095, 2021. 2, 3, 6, 7

Fine-tuning

Fine-tune VATT's
- vision transformer for vision-tasks
- audio transformer for audio-tasks

Zeo-shot:
- Feed video-text pairs
- Extract representation from common space
- Rank videos based on their similarities to the input text



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.

			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.

Learned Feature Visualization

Justification for design choice:
It is worth noting that there is no clear difference between the modality-agnostic features and the modality-specific ones

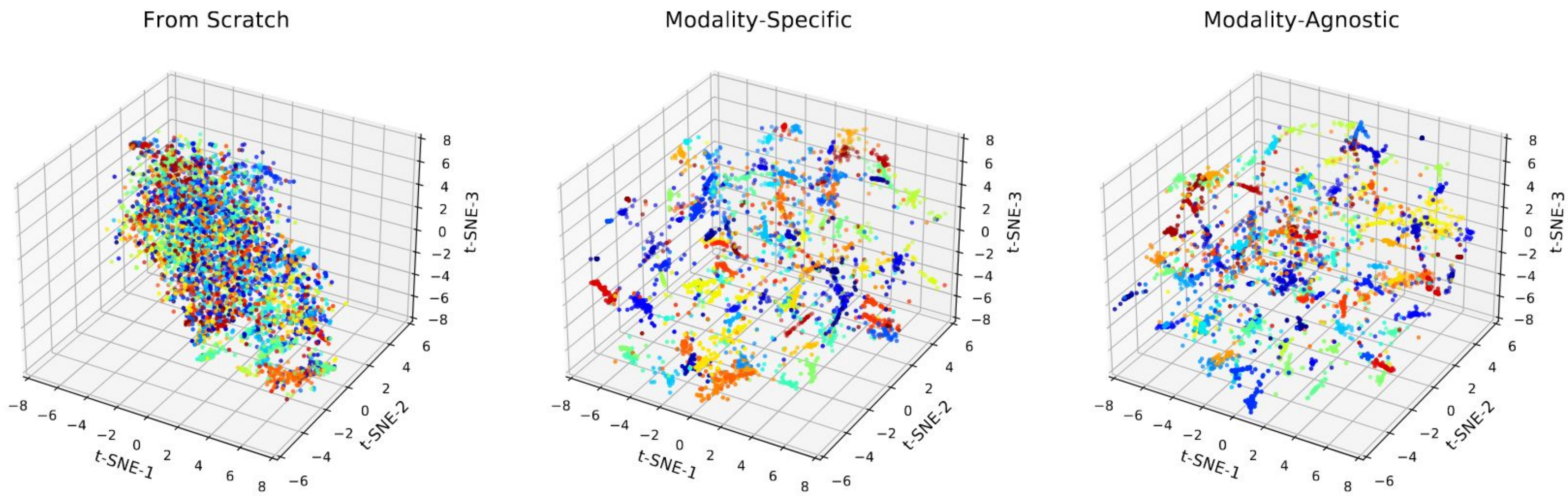
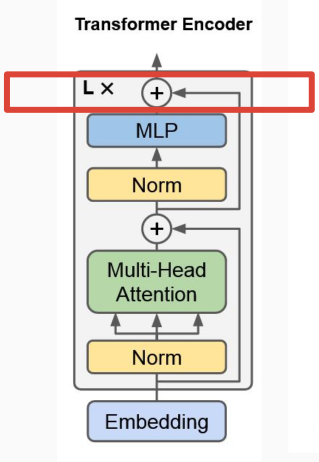


Figure 2: t-SNE visualization of the feature representations extracted by the vision Transformer in different training settings. For better visualization, we show 100 random classes from Kinetics-400.

Model Activations

Different layers/nodes have different jobs, depending on the modality:

- Early nodes for text
- Middle layer for video/audio
- Later layer for aggregation



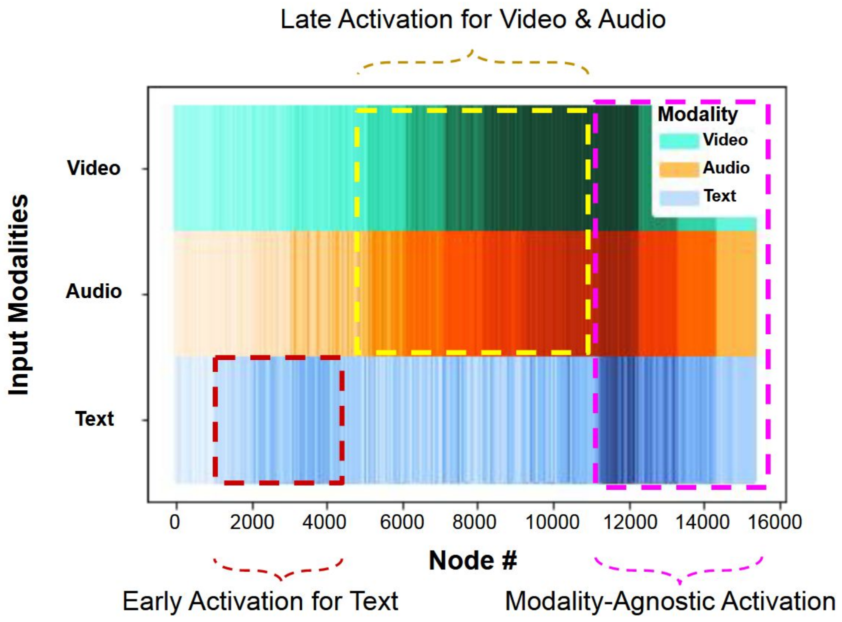


Figure 4: The average node activation across the Modality-Agnostic-Medium VATT while feeding a multimodal video-audio-text triplet to the model.

the average activation of each node at the output of the MLP module, before the residual addition

Drop-token Results

- Randomly drop 75%,50%, 25%, 0%
- Prefer High-resolution inputs

	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.

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.

Questions?