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

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

Submitted Apr 22, 2021; Last Revised Dec 7, 2021 (NeurIPS 2021)

Presented by : David K., Li Hui, Junjie

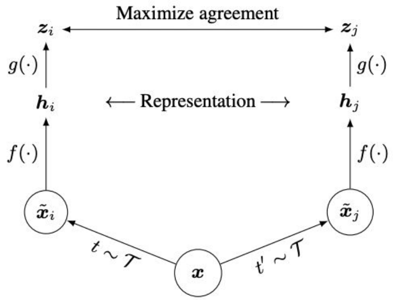
Research Questions & Motivation

- Can we use **ONE architecture** to learn vision, audio and language representations ?
- Can we share **ONE backbone** across all modalities ?
- How could we use **RAW inputs** with this model ?
- How could we **drop redundancy** in raw inputs ?
- Can we train this pipeline **without supervision** ?

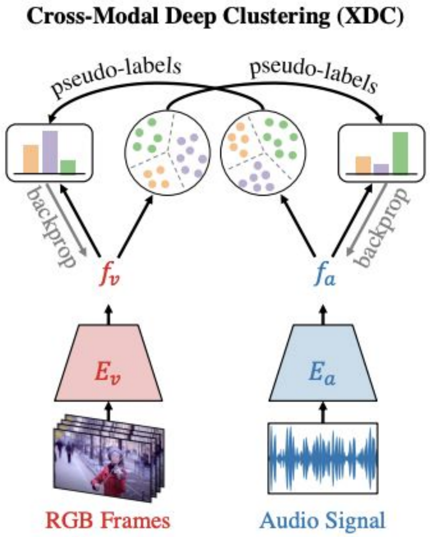
Adapted from : Google Research Presentation

Related Work - Self-Supervised Learning

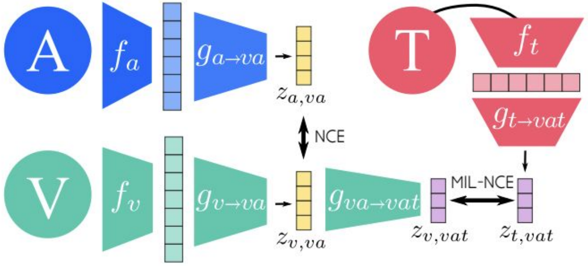
SimCLR : Chen et al, 2020 **(Vision)**



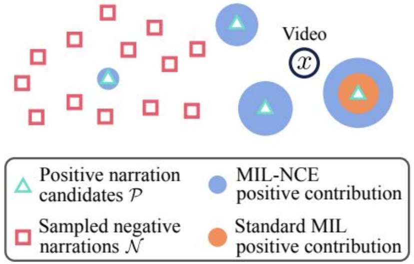
XDC : Alwassel et al, 2020
(Vision + Audio)



MMV : Alayrac, 2020
(Vision + Audio + Text)

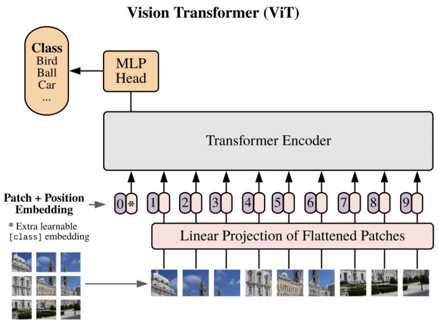


MIL-NCE : Miech et al, 2020
(Vision + Text)

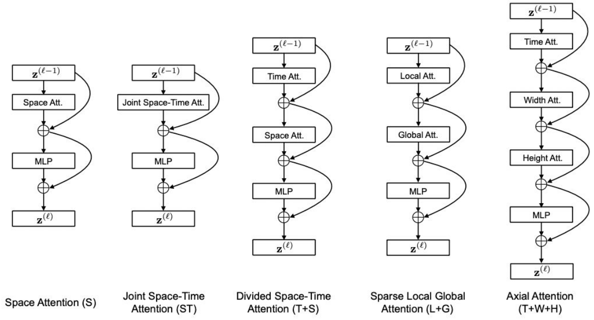


Related Work - Transformers

ViT : et al, 2020 **(Image)**

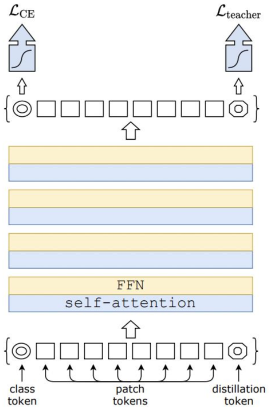


TimeSformer : Bertasius et al, 2021 **(Video)**

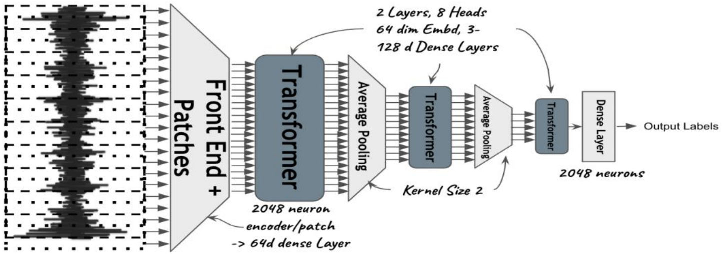


Multimodal ?

DeiT : et al, 2020 **(Image)**



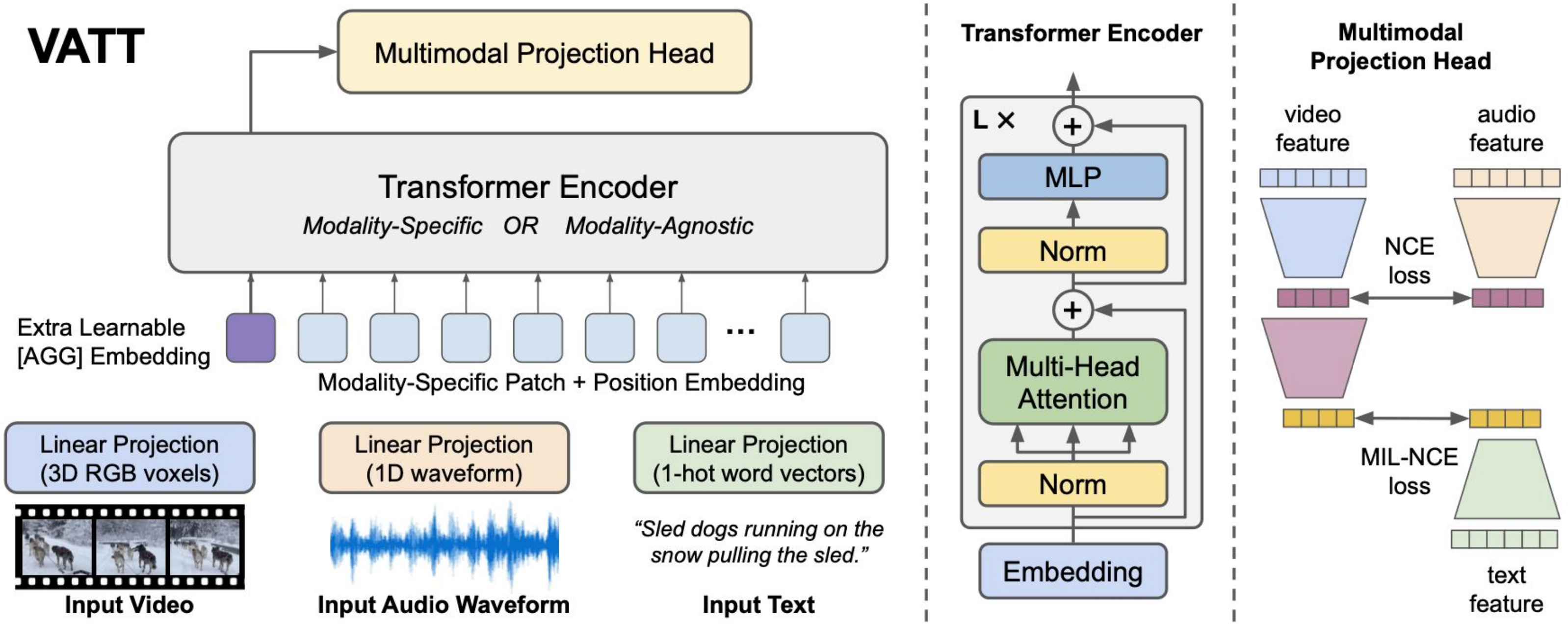
Audio Transformer : Verma, Berger 2021 **(Audio)**



Critiques on Prior Methods

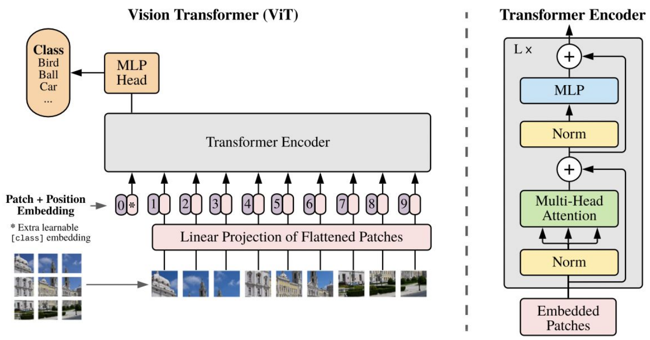
- **Large-scale supervised training** of Transformers :
  - Extremely costly and time-consuming for manual labelling of data
  - In reality, many visual data are unlabelled and unstructured
- **Ad-hoc network design** on previous works
  - Separate weights on different modalities (no weight sharing)
  - Computationally intensive architecture
- **Low resolution input**
  - Limited performance

Introducing VATT



Architecture

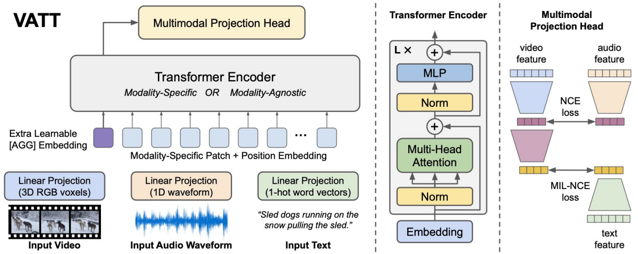
Overview of the VATT Architecture

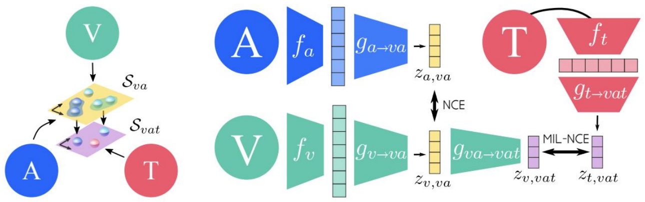


ViT

+

=





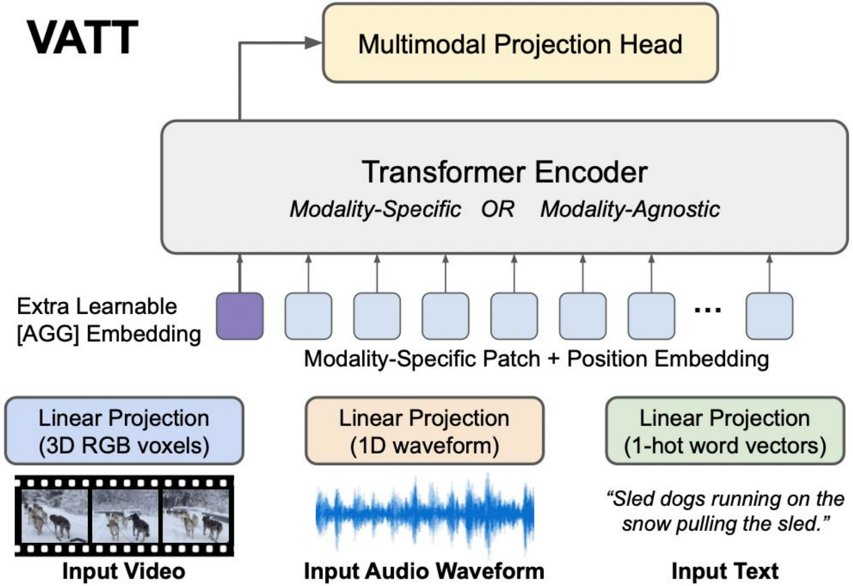
(c) **FAC**

(d) **FAC details**

MMV

VATT

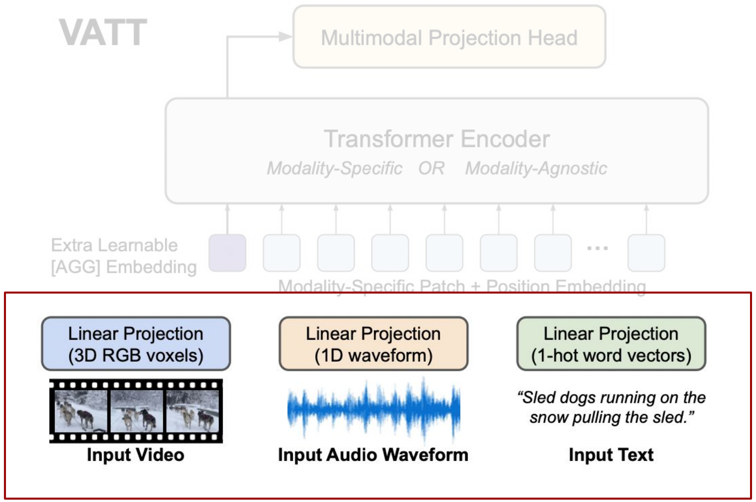
Overview of the VATT Architecture

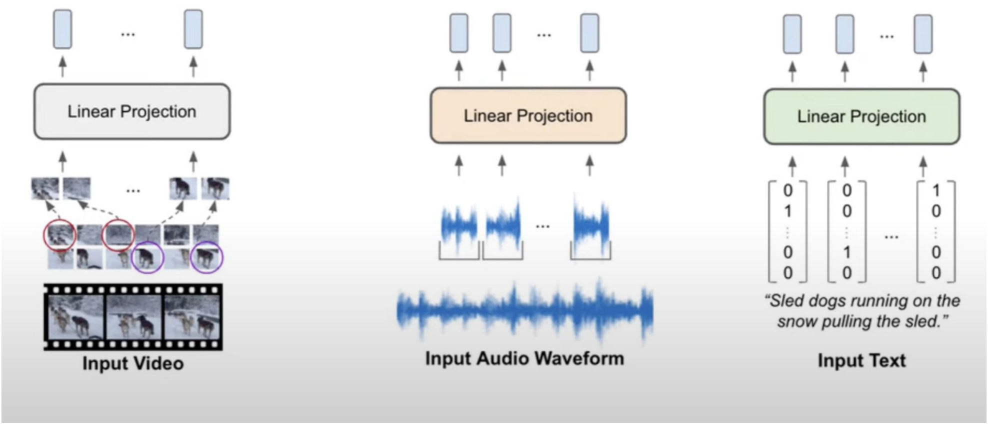


- Tokenization
- Positional Encoding
- DropToken
- Common Space Projection (MMV)
  - Self-Supervised Multimodal Versatile Network (NeurIPS'20)
- Multimodal Contrastive Learning
  - End-to-End Learning of Visual Representation from Uncurated Instructional Videos (CVPR'20)

Tokenization Layer

- A **modality-specific tokenization layer** that takes **raw signal as inputs** and **returns a sequence of vectors** to be fed to the Transformers

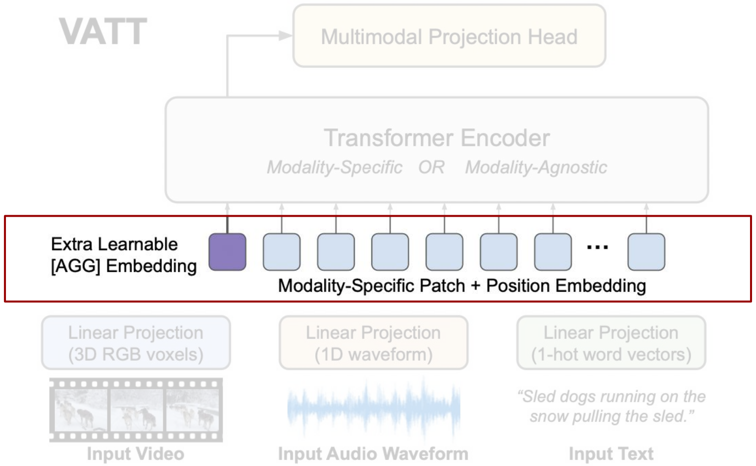


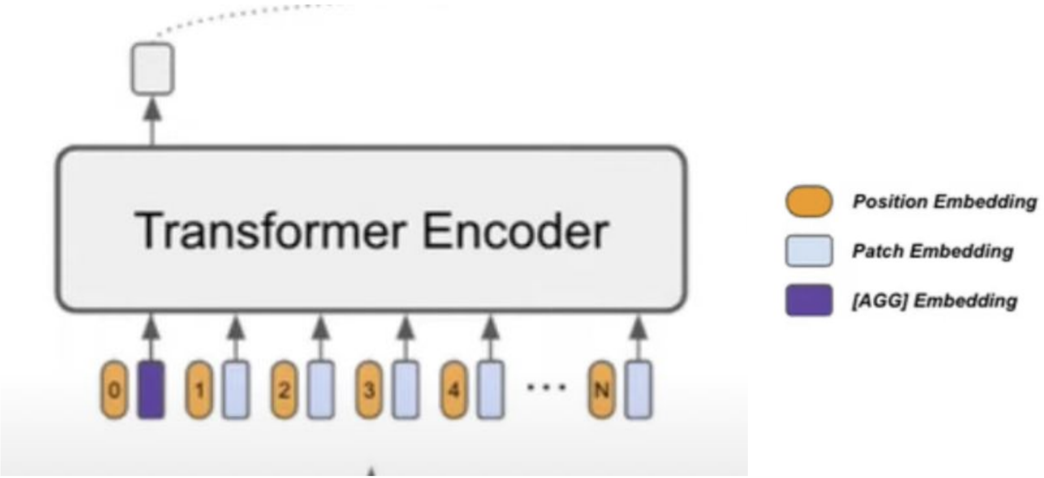


Patching (similar to ViT)

Positional Encoding

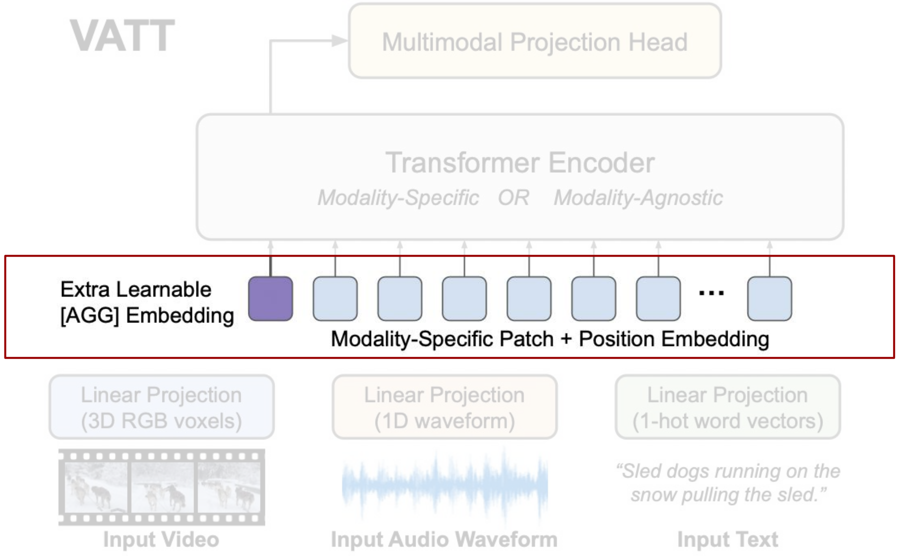
- Similar to standard ViT, **each modality has its own positional encoding**, which injects the **order** of tokens into the Transformer

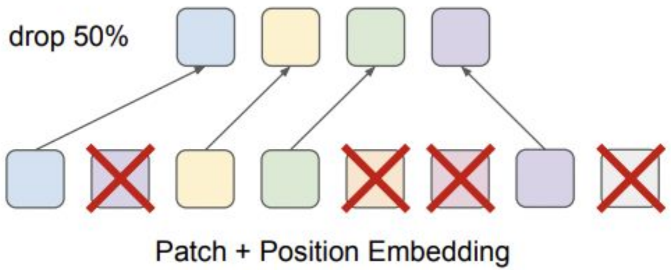




DropToken (video & audio inputs)

- Randomly sampled 50%* of the tokens and feed the sampled sequence to the Transformer
- Raw inputs : Low-resolution vs **High-fidelity + DropToken**





DropToken reduces computational complexity during *training*

*Good trade-off between accuracy and computational costs

Transformer Encoder

- Standard ViT architecture is used

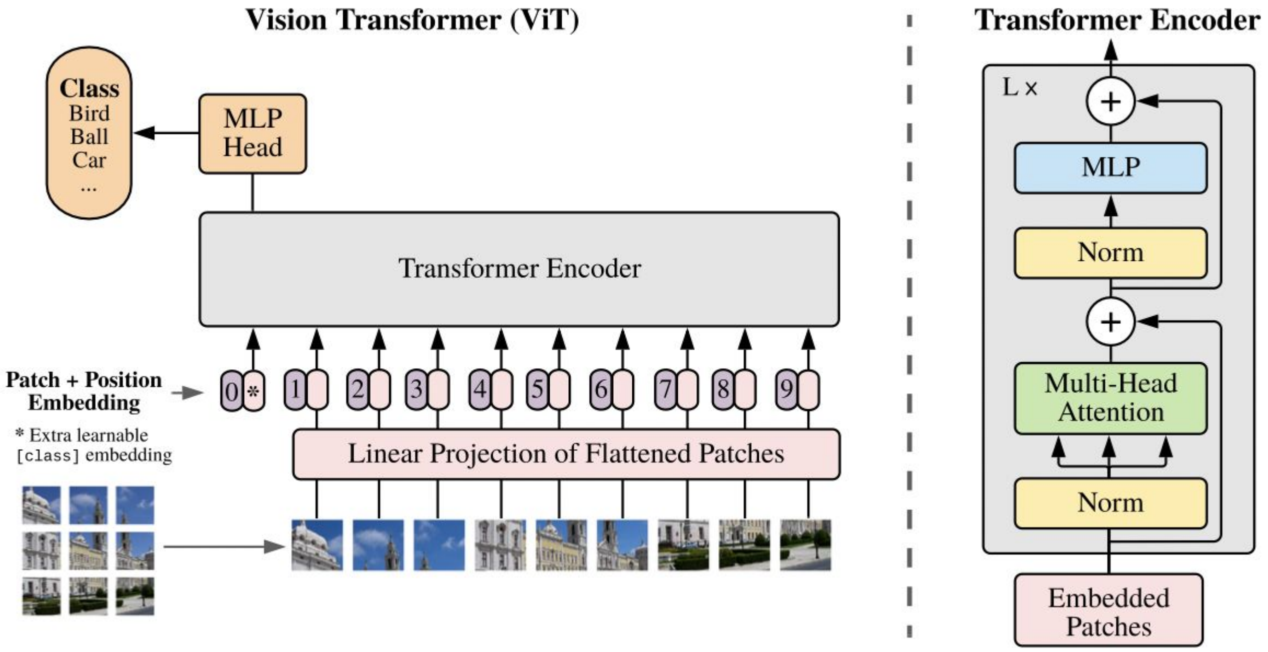
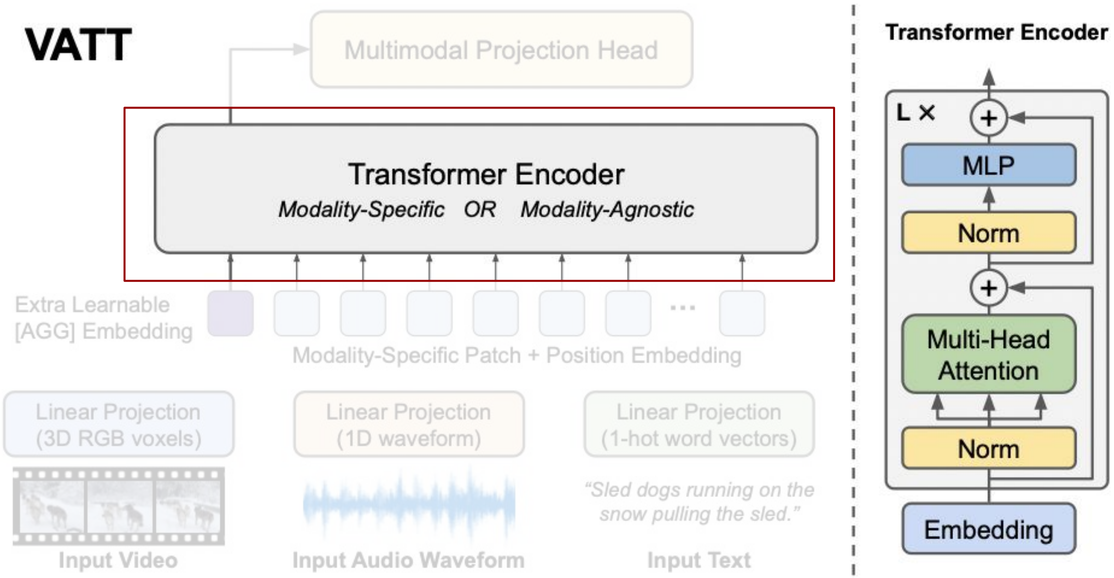


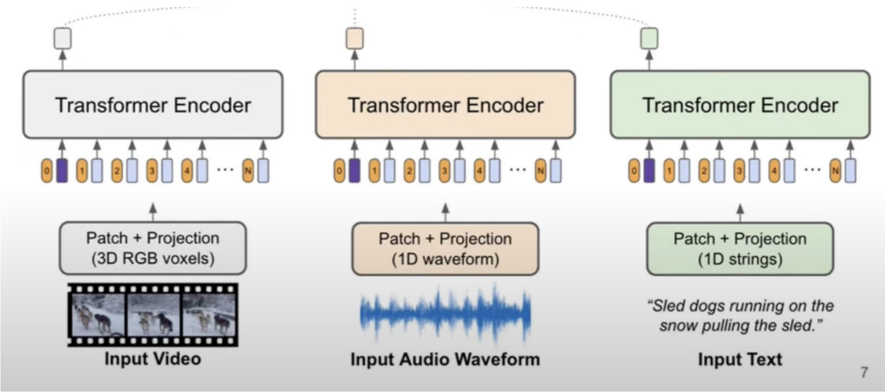
Figure from [An Image Is Worth 16X16 Words : Transformers for Image Recognition]() (ICLR'21)

Transformer Encoder

Two Settings of Transformer Encoder : **Modality-Specific** or **Modality-Agnostic**

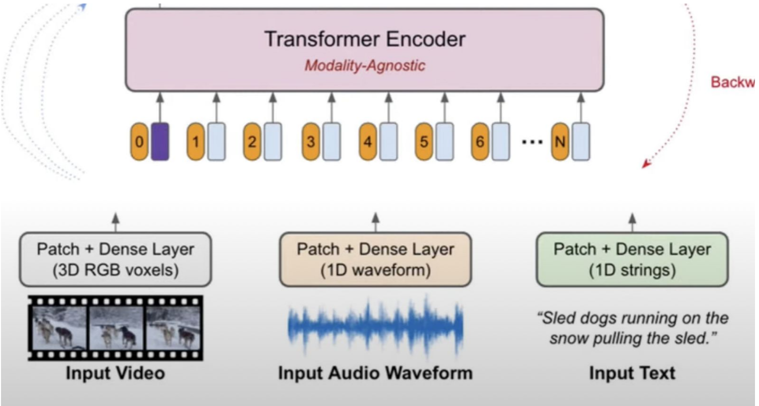


Modality-Specific vs Modality-Agnostic



(Left) **Modality-Specific** : the backbone Transformers are separate and have **specific weights for each modality**

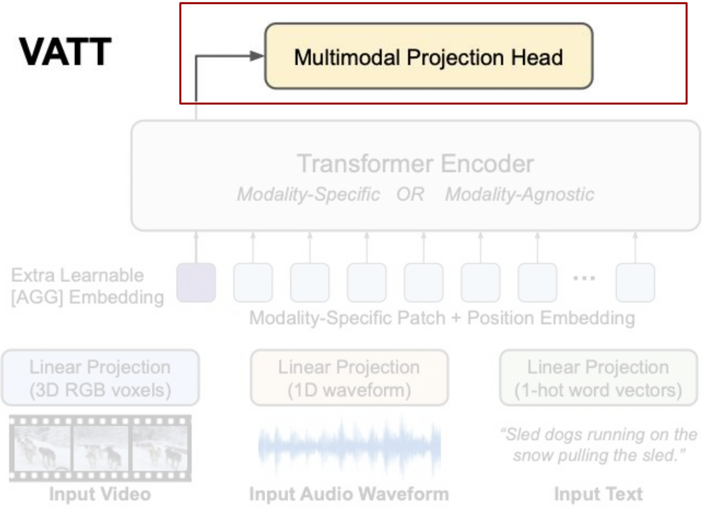
(Right) **Modality-Agnostic** : One single backbone Transformer and **weights are shared across all modalities**

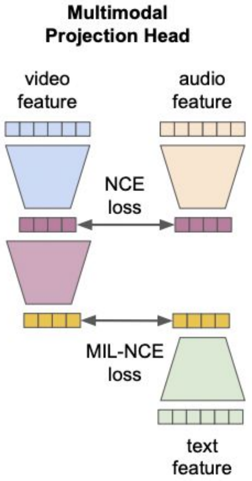


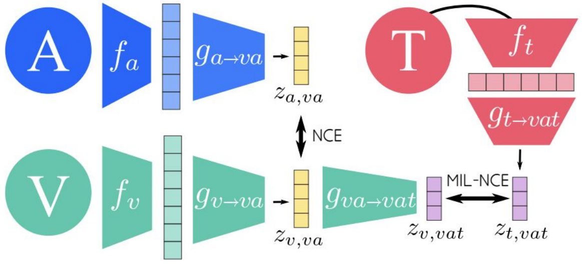
Adapted from : [Google Research Presentation]()

Architecture : Common Space Projection

- **Intuition : Different modalities have different levels of semantic granularity**
- We need **a common space** that is **semantically hierarchical** to “fuse” the multi-modality features



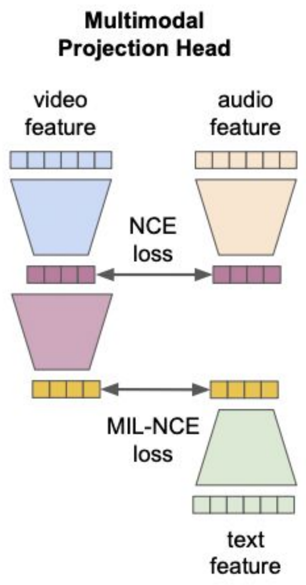


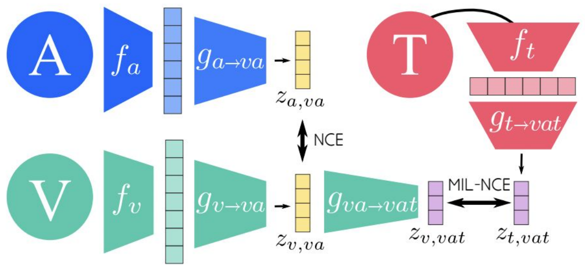


(d) **FAC details**

Rightmost figure from [Self-Supervised Multimodal Versatile Network](#) (NeurIPS’20)

Architecture : Multimodal Contrastive Learning

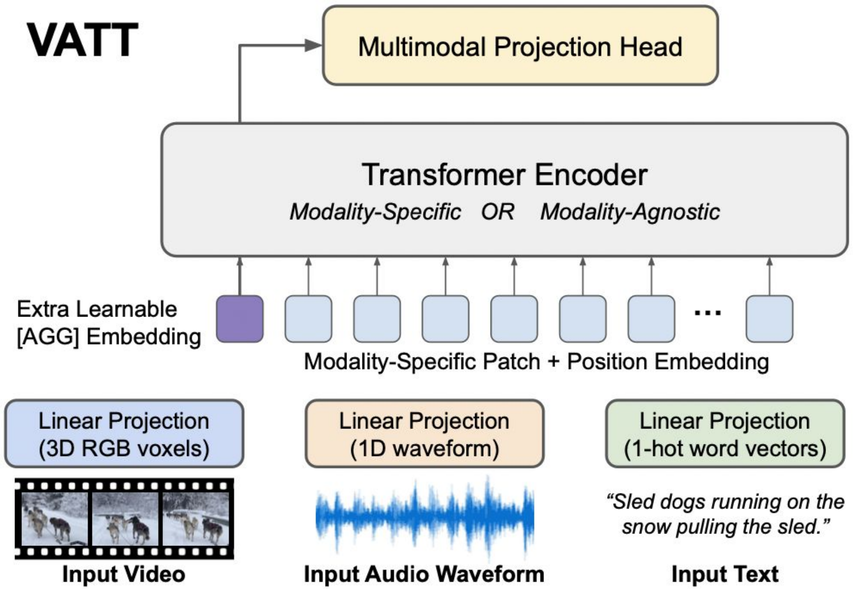




(d) **FAC details**

- Contrastive representation in self-supervised learning
- To align the pairs :
  - **Vision & Audio Pair** : Noise Contrastive Estimation (NCE) loss
  - **Vision & Text Pair** : Multiple-Instance-Learning-NCE (MIL-NCE) loss

VATT Architecture Recap



- Tokenization
- Positional Encoding
- DropToken
- Common Space Project
- Multimodal Contrastive Learning

Experiments

Experimental Setup

Pre-trained on HowTo100M and AudioSet and evaluated on:

1. **Video Action Recognition** on UCF101, Kinetics-400/600, and Moments in Time
2. **Audio Event Classification** on ESC50 and AudioSet
3. **Zero-shot Video Retrieval** on YouCook2 and MSR-VTT
4. **Image Classification** on ImageNet

VATT-MA-Medium → Modality-agnostic

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.

Achieves SOTA and set a new record at the time on the biggest benchmarks.

Outperforms TimeSFormer, a ViT-inspired fully-supervised approach.

Modality-agnostic VATT is competitive with:

- Base model
- fully-supervised

Audio Event Classification

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.

Outperforms CNN-based approaches to audio event classification

Modality-agnostic model is competitive with modality-specific model.

Image 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.

Pre-trained on video data but achieves competitive results vs. fully-supervised training on an immense image dataset.

Zero-Shot Video Retrieval

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

Does not set any records but is competitive with other self-supervised models.

Thank You