Indian Institute of Technology, Guwahati

DA323 Course Project Presentation

Exploring VATT: Transformers for Multimodal Learning (Audio, Video, and Text)

08 May, 2025

Outlook ID: raunit@iitg.ac.in

Overview

Motivation

Avoids supervised pre-training costs and biases.

Processes raw, unlabeled video, audio, text data.

Matches or beats CNNs in vision/audio tasks.

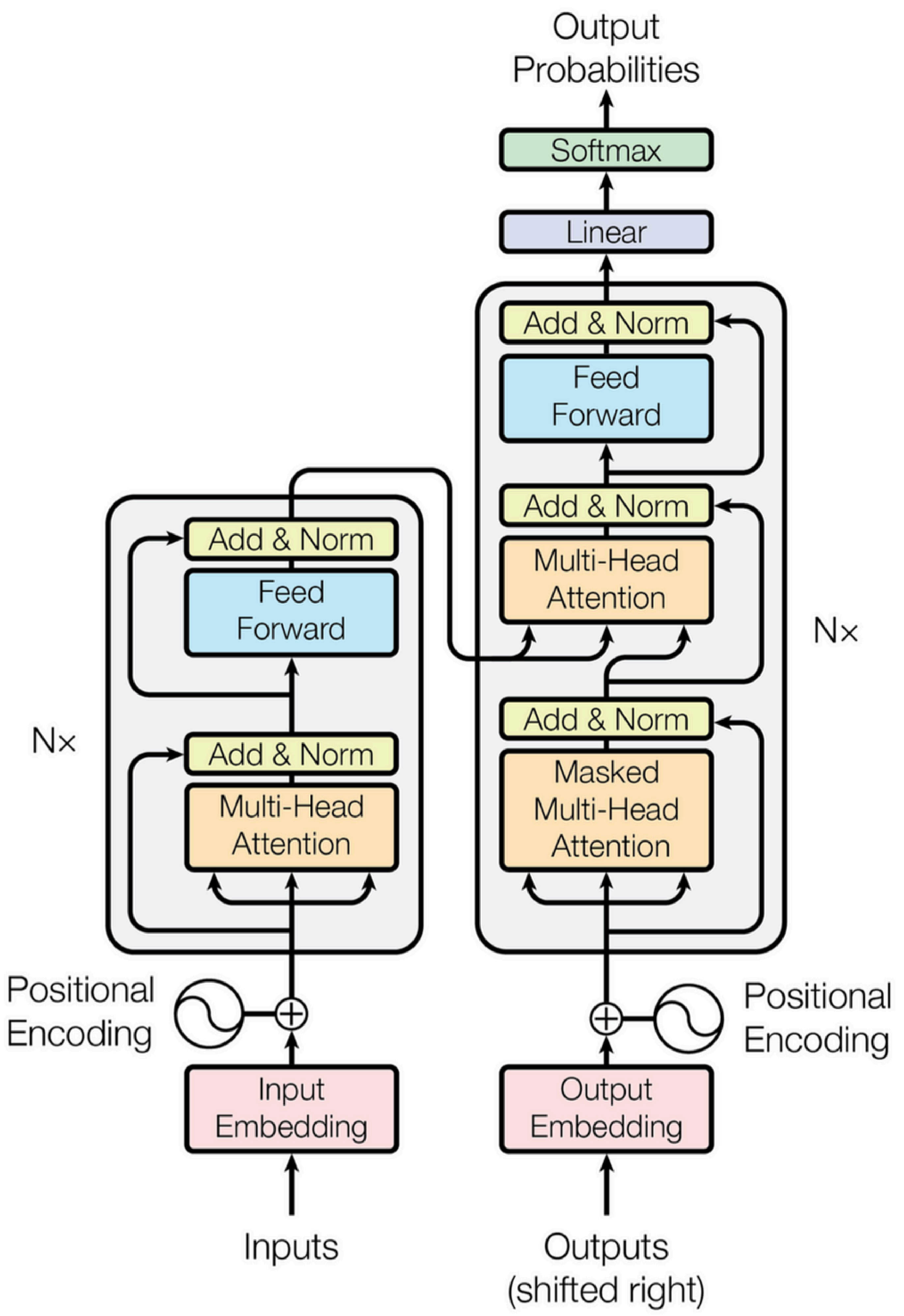


Figure 1: The Transformer - model architecture.

Historical Perspective

From CNNs to Transformers: A Historical Shift

Early AI: CNNs dominated vision/audio, LSTMs for text (pre-2017).

2017 Breakthrough: Transformers introduced via "Attention is All You Need", revolutionizing NLP with self-attention.

NLP Success: BERT and GPT set records with large-scale pre-training.

Vision Transition: Vision Transformer (ViT) matched CNNs in image tasks using pure Transformers (2020).

Historical Perspective

Rise of Multimodal Learning

Transformer Impact: Unified models like VideoBERT, ViLBERT, LXMERT, CLIP advanced cross-modal learning.

VATT (2021): Convolution-free, self-supervised learning on raw video, audio, text; achieves 82.1% on Kinetics-400.

Shift: From modality-specific to general-purpose, modality-agnostic Transformers without labeled data.

Key Learnings

Power of Convolution-Free Transformers Across Modalities

Effectiveness of convolution-free Transformer architectures in processing raw multimodal data-video (RGB frames), audio (waveforms), and text (transcripts)-without relying on Convolutional Neural Networks (CNNs).

Self-Supervised Learning with Multimodal Contrastive Losses

By training on unlabeled datasets like HowTo100M and AudioSet, VATT aligns representations across modalities in a common semantic space, capturing cross-modal relationships without human-annotated labels.

Modality-Agnostic Transformer Potential

A fascinating aspect of VATT is the exploration of a single-backbone, modality-agnostic Transformer where weights are shared across video, audio, and text modalities (with separate tokenization and projection layers).

Key Learnings

DropToken for Computational Efficiency

VATT introduces DropToken, a novel technique to manage the quadratic computational complexity of Transformers ($O(N^2)$ with respect to input tokens).

Transferability Across Domains

Despite being pre-trained on multimodal video data, fine-tuning VATT's vision Transformer on ImageNet yields a top-1 accuracy of 78.7%, close to ViT's 79.9% and far surpassing the 64.7% achieved by training the same Transformer from scratch.

Approach

Two Major Settings

The backbone Transformers are separate and have specific weights for each modality.

The Transformers share weights, namely, there is a single backbone Transformer applied to any of the modalities.

Approach: Tokenisation And Positional Encoding

VATT processes raw signals from video, audio, and text into vector sequences for Transformers using modality-specific tokenization and positional encoding.

Video	Audio	Text
Raw RGB video clips (T×H×W) are split into patches of t×h×w×3 voxels.	Raw waveforms of length T′ are divided into ⌈T′/t′⌉ segments of t′ amplitudes, projected to d-dimensional vectors with weight Wap∈Rt′×d.	Word sequences are mapped to v-dimensional one-hot vectors based on a vocabulary of size v, then projected to d-dimensional embeddings with weight Wtp∈Rv×d.

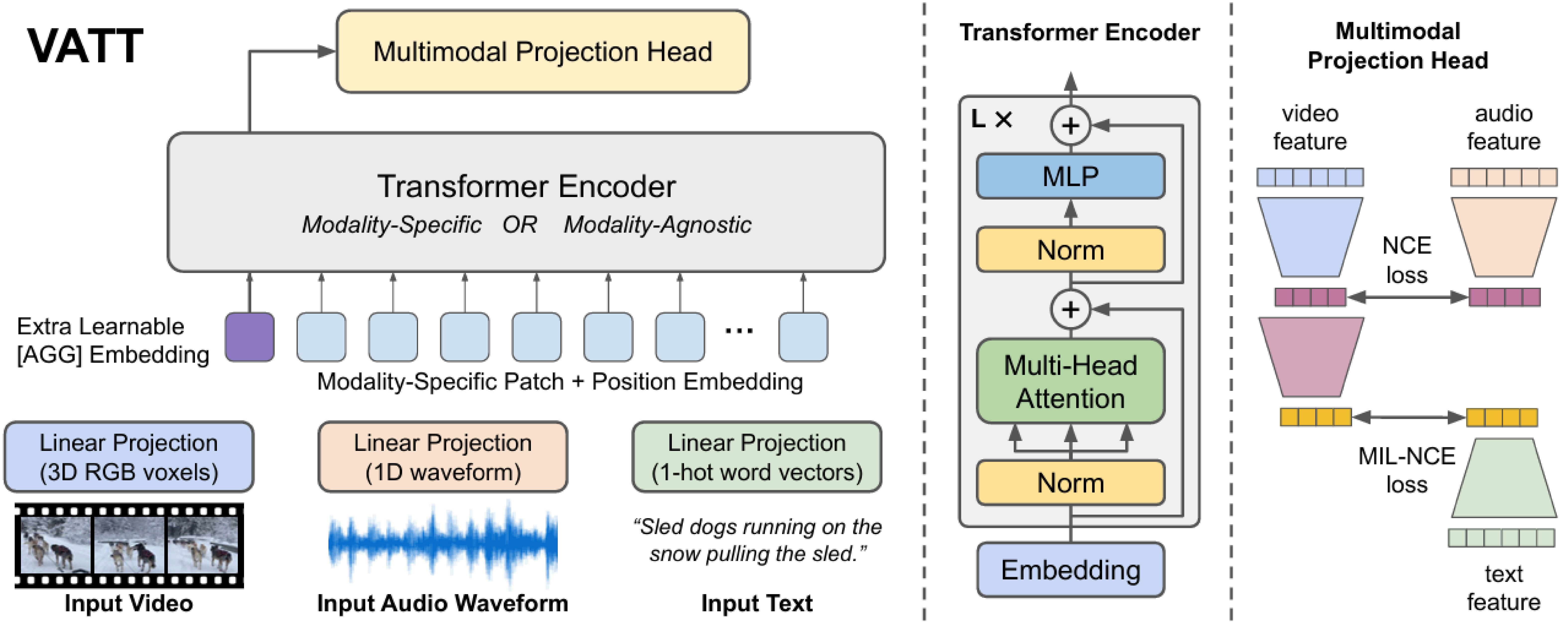
Approach: Tokenisation And Positional Encoding

DROPTOKEN

After tokenizing video and audio inputs into sequences, DropToken randomly samples a portion of these tokens, feeding only the sampled subset-rather than the full set-to the Transformer.

This is critical because a Transformer's computational complexity is quadratic, O(N2), where N is the number of input tokens.

Approach: Transformer Architecture



Approach: Common Space Projection

Semantically Hierarchical Common Space Mapping

$$\mathbf{z}_{v,va} = g_{v\to va}(\mathbf{z}_{\text{out}}^{\text{video}}), \mathbf{z}_{a,va} = g_{a\to va}(\mathbf{z}_{\text{out}}^{\text{audio}})$$

$$\mathbf{z}_{t,vt} = g_{t\to vt}(\mathbf{z}_{\text{out}}^{\text{text}}), \mathbf{z}_{v,vt} = g_{v\to vt}(\mathbf{z}_{v,va})$$

Projection Head Design

For ga→va, gt→vt, and gv→vt, a linear projection is used. For gv→va, a two-layer projection with ReLU activation between layers is applied to capture more complex transformations.

Approach: Multimodal Contrastive Learning

Noise Contrastive Estimation (NCE) for Video-Audio Pairs

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

ProjectioMultiple Instance Learning NCE (MIL-NCE) for Video-Text Pairsn Head Design.

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

Overall Loss Objective

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

Experiments: Pretraining Setup

All pre-training experiments apply DropToken with a 50% drop rate.

Training employs the Adam optimizer with a quarter-period cosine learning rate schedule from 1e-4 to 5e-5, 10k warmup steps, and 500k total steps with a batch size of 2048.

Common space projections use dimensions dva=512 for video-audio space Sva and dvt=256 for video-text space Svt, with a temperature τ=0.07 and loss weight λ=1 in the combined loss equation.

VATT is pre-trained on a combination of AudioSet and a subset of HowTo100M datasets, adhering to YouTube's policies.

Experiments: Downstream Evaluation

Pre-trained VATT models are evaluated on four major downstream tasks across 10 datasets:

Video Action Recognition: UCF101, HMDB51, Kinetics-400, Kinetics-600, and Moments in Time

Audio Event Classification: ESC50 and AudioSet.

Zero-Shot Text-to-Video Retrieval: YouCook2 and MSR-VTT, assessing video-text common space quality.

Image Classification: ImageNet, testing vision backbone transferability.

Results: 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

Results: Fine-tuning for 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

Results: Fine-tuning for Image Classification & Zero-shot Text-to-Video Retrieval.

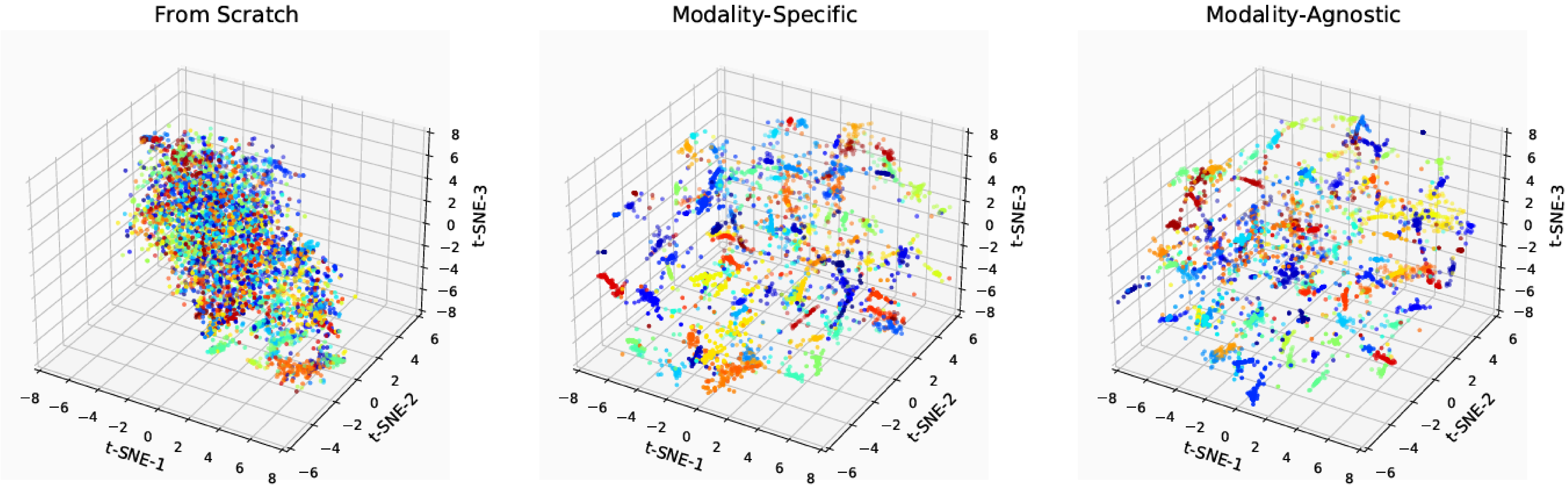
METHOD	PRE-TRAINING DATA	TOP-1	TOP-5
iGPT-L [16]	ImageNet	72.6	-
ViT-Base [25]	JFT	**79.9**	-
VATT-Base	-	64.7	83.9
VATT-Base	HowTo100M	78.7	93.9

Table 3: Fine tuning 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.

Results: Feature Visualization



Results: Effect of DropToken

	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.

What Surprised Me?

Exceptional Transferability

Modality-Agnostic Success

Raw Data Mastery

CNN Parity

Scope of Improvement

Text Data Quality

Computational Efficiency

Broader Modalities

Mixture-of-Experts

Indian Institute of Technology, Guwahati

Thank You

23 April, 2025