“Self-Supervised MultiModal Versatile Networks”

Sushodhan Sudhir Vaishampayan (246150101)

Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology, Guwahati

May 8, 2025

Agenda

About the paper

- **Title** - Self-Supervised MutliModal Versatile Networks
- **Published** - Advances in Neural Information Processing Systems 33 (NeurIPS 2020)
- **Authors** - Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, Andrew Zisserman

Multimodal

Definition

"Multi" - Many

"Modality" - A particular form of sensory perception

"MultiModality" - "Multi" + "Modality". From many forms of sensory perceptions

Example

To understand a person's emotion we take input their expression and body language (visual) along with the variations in the person's pitch (audio).

Embeddings

Definition

Transforming real life objects like text, image, audio into a mathematical form (usually vectors) to be understood by computers.

- **Motivation** - How to compare the word 'Man' to 'Vehicle' extracted from textual data?
- **Solution** - Convert the words (objects) to real-valued tensors (usually 1D).
- **Why?** - Computers can compare scalars, vectors, matrices, and n-d tensors.

Self-Supervised Learning

Definition

Given unlabeled data, Self-Supervised Learning(SSL) aims to leverage the **inherent internal structure and relationships between different parts of the data**.

Pretext Tasks

Definition

SSL takes place by solving tasks, called as "Pretext Tasks" for which 'ground truth' can be extracted from the data itself.

Example

- Jigsaw puzzle- Given scrambled patches of images (unlabeled), we learn a model to rearrange the patches correctly.
- Masked Language Modeling (MLM) - Given a sentence, predicted the missing words (masked intentionally) from the sentence.

Downstream Tasks

Definition

The actual tasks that we intend to solve from the representations learned during *Pretext tasks*.

Example

- Image Classification
- Sentiment Analysis

Objective

To learn a *MultiModal Versatile* network that

1. Takes input from any of the **three modalities** (visual, audio, and text)
2. Consider the **specificity** of the modalilties (fine or coarse grained)
3. Allow **easy comparison** between different modalities
4. Should be applicable to visual data both in the form of **dynamic videos** and **static images**.

Self-Supervised Learning from Single Modality

- Predicting relative position of patches[1][2]
- Colorization[3]
- Predicting Orientation[4]
- Invariance to Transformation[5][6]

Other Related Work

- Vision and Language[7][8]
- Vision and Audio (predict whether visual and audio signals belong to the same video)[9]
- Vision, Audio, and Language[10]
- From Video to Image[11]

Notations

Input

- Video $x \in \chi$
- Modality $M : x \rightarrow x_m, m \in M$
- Vision $x_v \in \chi_v$ - Few second sequence of RGB frames
- Audio $x_a \in \chi_a$ - 1D audio sample
- Text $x_t \in \chi_t$ - discrete word tokens

Notations

Representaions

- $f_m : \chi_m \to \mathbb{R}^{d_m}$ - modalilty specific backbone neural network.
- $S_s \subset \mathbb{R}^{d_s}$ - shared subspace for comparison[1]
- $g_{m\to s} : \mathbb{R}^{d_m} \to \mathbb{R}^{d_s}$ - projection head to space S_s
- $z_{m,s} = g_{m\to s}(f_m(x_m))$ - vector representing input modality x_m in space S_s

1

- $s = va$, then S_{va} is joint visual and audio space
- $s = vat$, then S_{vat} is joint visual, audio, text space

Multimodal Versatile Networks

- Option I: Shared Space - $S_{vat} \subset \mathbb{R}^{d_s} \Rightarrow z_{v,vat}$
- Option II: Disjoint Spaces - S_{va} and $S_{vt} \Rightarrow z_{v,va} \neq z_{v,vt}$
- Option III: Fine And Coarse Spaces (FAC) - S_{va} and S_{vat}

Multimodal Versatile Networks

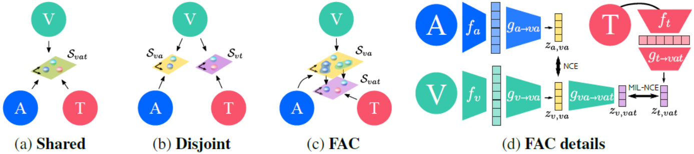

Figure 1: (a)-(c) Modality Embedding Graphs, (d) Projection heads and losses for the FAC graph. V=Vision, A=Audio, T=Text.

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

Multimodal Versatile Networks

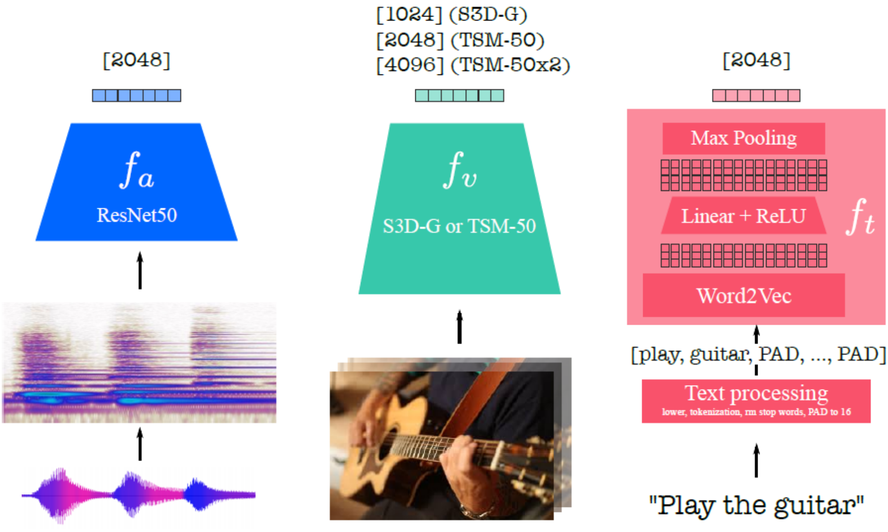

Figure: Backbone architecture for audio, vision and text.

Multimodal Contrastive Loss

$$\mathcal{L}(x) = \lambda_{va} NCE(x_v, x_a) + \lambda_{vt} MIL\text{-}NCE(x_v, x_t)$$

$$NCE(x_v, x_a) = -log\left(\frac{exp(z_{v,va}^T z_{a,va}/\tau)}{exp(z_{v,va}^T z_{a,va}/\tau) + \sum_{z' \sim \mathcal{N}(x)} exp(z'^T_{v,va} z'_{a,va}/\tau)}\right)$$

Video to Image Network Deflation

Aim

The output of the deflated video network on a single image must be identical to the output of the single-image static-video for the same image using original video network

- For 3D Convolutional Networks - Sum the 3D Spatiotemporal filters over temporal dimension to obtain 2D filters
- For TSM networks, turn of the channel shifting.

Network Architectures and Hyperparameters

Video

- S3DG[12]($d_v = 1024$)
- TSM[13] with ResNet50 backbone($d_v = 2048$)
- TSM with ResNet50$\times$2[14] backbone($d_v = 4096$)

Audio

- ResNet50($d_a = 2048$)

Text

- Word2Vec[15]($d_t = 2048$)

Network Architectures and Hyperparameters

Hyperparameters

- $S_{va} \in \mathbb{R}^{512}$
- $S_{vat} \in \mathbb{R}^{256}$
- $\tau = 0.07$
- $\textit{initial learning rate} = 0.002(\text{Adam}[16])$

Datasets

Type	Dataset
Training[2]	HowTo100M + AudioSet
Testing	UCF101
	HMDB51
	Kinetics600
	ESC-50
	AudioSet
	YouCook2
	MSRVTT
	PASCAL
	Imagenet

Table: Datasets

[2] Instances from same sample are positive; from different instances are negative

Preprocessing of Inputs

Video

- Temporal sampling (16/32 frames subclip per video)
- Resizing (minimum side to 224)
- Extract random crop (200×200)
- Scale Jiterring (width(or height)$*s \sim Unif(0.8, 1.2)$)
- Horizontal Flipping
- Color Augmentation(brightness, saturation, contrast, hue)
- RGB clipped in $[0.0, 1.0]$

Preprocessing of Inputs

Audio

- log MEL Spectrogram with 80 bins
- 2 second audio ingestion

Text

- Remove stop words
- Retain maximum or padding to 16 words
- Extract 300-dimensional Google News pre-trained word2vec

Downstream Tasks

Task	Dataset
Linear classifier	UCF101/HMDB51
Fine Tuning	UCF101/HMDB51
Linear classifier	Kinetics600
Linear classifier	ESC-50
Linear classifier	AudioSet
Zero-shot text-to-video retrieval	YouCook2/MSRVTT
Linear classifier	PASCAL/ImageNet

Table: Downstream Tasks

Results

Method	f_v (#params)	Train data	years	Mod.	UCF101 Linear	UCF101 FT	HMDB51 Linear	HMDB51 FT	ESC-50 Linear	AS MLP	K600 Linear
MIL-NCE [49]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [49]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [41]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [41]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [32]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [67]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [4]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [4]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [64]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [55]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	**89.2**		
GDT [62]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [62]	R(2+1)D-18 (33.3M)	IG65M	21	VA		**95.2**		72.8			
VA only (ours)	R(2+1)D-18 (33.3M)	AS	1	VA	83.9	91.5	60.0	70.1	85.6	29.7	55.5
VA only (ours)	S3D-G (9.1M)	AS	1	VA	84.7	90.1	60.4	68.2	86.1	29.7	59.8
VA only (ours)	S3D-G (9.1M)	AS+HT	16	VA	86.2	91.1	61.5	68.3	87.2	30.6	59.8
MMV FAC (ours)	S3D-G (9.1M)	AS+HT	16	VAT	89.6	92.5	62.6	69.6	87.7	30.3	68.0
MMV FAC (ours)	TSM-50 (23.5M)	AS+HT	16	VAT	91.5	94.9	66.7	73.2	86.4	30.6	67.8
MMV FAC (ours)	TSM-50x2 (93.9M)	AS+HT	16	VAT	**91.8**	**95.2**	**67.1**	**75.0**	88.9	**30.9**	**70.5**
Supervised [19, 40, 64, 71, 87]						96.8	71.5	75.9	86.5†	43.9	81.8

Figure: Comparison of learnt representations versus the state-of-the-art. Top-1 accuracy is reported for UCF101, HMDB51, ESC-50, kinetics600 and mAP for AudioSet.

Results

Method	V→I	Train data	PASCAL(mAP)	ImageNet(top1)	ImageNet(top5)
Supervised S3D-G	def	Kinetics	67.9	42.8	68.0
MMV S3D-G	n-def	AS+HT	41.8	20.7	40.5
MMV S3D-G	def	AS+HT	71.4	45.2	71.3
MMV S3D-G	i-inf	AS+HT	72.1	46.7	72.5
Supervised TSM	def	Kinetics	66.9	43.4	68.3
MMV TSM	n-def	AS+HT	34.4	10.9	24.6
MMV TSM	def	AS+HT	74.8	50.4	76.0
MMV TSM	i-inf	AS+HT	75.7	51.5	77.3
Supervised TSMx2	def	Kinetics	66.9	47.8	72.7
MMV TSMx2	n-def	AS+HT	45.6	20.3	39.9
MMV TSMx2	def	AS+HT	77.4	56.6	81.4
MMV TSMx2	i-inf	AS+HT	77.4	57.4	81.7
SimCLR ResNet50	/	ImageNet	80.5	69.3	89.0
SimCLR ResNet50x2	/	ImageNet	/	74.2	92.0
SimCLR ResNet50x4	/	ImageNet	84.2	76.5	93.2

Table: Image classification results on PASCAL and ImageNet. "V→I" denotes the image handling strategy for the video networks: naive deflation (no training of γ and β), deflation (proposed), and input-inflation (video net ingesting 32-frame static videos).

Conclusion

- The paper conducts self-supervised training to build *versatile* networks for *vision*, *audio* and *language*
- The trained network is tested on downstream task like *action and audio classification*
- It is also shown, how a network trained for videos can be used for images

My Take on the Paper

- The papers proposes an interesting idea of having a common space for embeddings of video, audio and language signals
- This common space can help for comparison between different modalities
- The granularity is also taken into account for constructing these embeddings
- The **comparison with state-of-the-art techniques doesn't seem fair enough** for downstream tasks as it is done on a very small subset of selective performance measures

Carl Doersch, Abhinav Gupta, and Alexei A Efros.
Unsupervised visual representation learning by context prediction.
In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis.
Unsupervised representation learning by predicting image rotations.
arXiv preprint arXiv:1803.07728, 2018.

Richard Zhang, Phillip Isola, and Alexei A Efros.
Colorful image colorization.
In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

Mehdi Noroozi and Paolo Favaro.
Unsupervised learning of visual representations by solving jigsaw puzzles.
In *European conference on computer vision*, pages 69–84. Springer, 2016.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox.
Discriminative unsupervised feature learning with convolutional neural networks.
Advances in neural information processing systems, 27, 2014.

Longlong Jing and Yingli Tian.
Self-supervised spatiotemporal feature learning by video geometric transformations.
arXiv preprint arXiv:1811.11387, 2(7):8, 2018.

A Frome, GS Corrado, J Shlens, et al.
A deep visual-semantic embedding model.
Proceedings of the Advances in Neural Information Processing Systems, pages 2121–2129.

Jason Weston, Samy Bengio, and Nicolas Usunier.
Wsabie: Scaling up to large vocabulary image annotation.
In *IJCAI*, volume 11, pages 2764–2770. Citeseer, 2011.

Relja Arandjelovic and Andrew Zisserman.

Look, listen and learn.
In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba.
See, hear, and read: Deep aligned representations.
arXiv preprint arXiv:1706.00932, 2017.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang.
Dual encoding for zero-example video retrieval.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9346–9355, 2019.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy.
Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification.
In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

Ji Lin, Chuang Gan, and Song Han.
Tsm: Temporal shift module for efficient video understanding.
In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer.
Revisiting self-supervised visual representation learning.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1920–1929, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.
Efficient estimation of word representations in vector space.
arXiv preprint arXiv:1301.3781, 2013.

Diederik P Kingma and Jimmy Ba.
Adam: A method for stochastic optimization.
arXiv preprint arXiv:1412.6980, 2014.