Self-Supervised MultiModal Versatile Networks

Jean-Baptiste Alayrac[1*] **Adrià Recasens**[1*] **Rosalia Schneider**[1*] **Relja Arandjelović**[1*]

Jason Ramapuram[2,3†] **Jeffrey De Fauw**[1] **Lucas Smaira**[1] **Sander Dieleman**[1]

Andrew Zisserman[1,4]

Presented by Pierre-Nicolas Perrin, Shoubin Yu - 11/02/2022

Motivation

- Humans experience the world in a multimodal way. Can we leverage as many modalities as possible in order to learn useful representations?

- Can we leverage audio, language, and vision to get good representations without labels?

Related Work

- Vision work inspired using contrastive loss and non-linear projection heads.
- Combining vision and language has been a topic of interest in the community, particularly using datasets with no manual labels.
- Vision and audio combination has been a little less explored.
    - Explore the co-occurence between visual and audio modalities in a video to learn good representations.
- Some past work focused on using audio, vision, and language to learn representations.
    - Harwath et al. use a dataset of images and audio descriptions to associate spoken words and their visual representation.
    - Aytar et al. train a cross-modal network with image, audio, and text modalities but rely on curated annotations whereas our paper does not.

Model

Multimodal versatile network with 4 properties**:**

1. It should be able to take as input any of the three modalities (visual, audio, language).
2. It should respect the specificity of the modalities (i.e audio and visual modalities are more fine-grained than language).
3. It should enable the different modalities to be easily compared even if they are never seen together during training.
4. It should be efficiently applicable to visual data coming in the form of dynamic videos or static images.

Model

- Given a set of unlabelled videos containing multiple modalities (RGB stream, audio track, linguistic narration) we wish to learn a model that has versatile properties as described in the previous slide.
- Focus on integrating three modalities: vision, audio, and text.
- Given a training set of N videos, the authors seek to learn the modality-specific representations as well as a way to compare streams across modalities.
    - Use a modality-specific backbone neural network to embed an instance from a specific modality into a common shared space R^d via a non-linear projection head.
    - Shared embedding space allows the comparison of different modalities by simple dot-product.

$f_m : \mathcal{X}_m \rightarrow \mathbb{R}^{d_m}$

Modality-specific backbone NN

$g_{m \rightarrow s} : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{d_s}$

Non-linear projection head

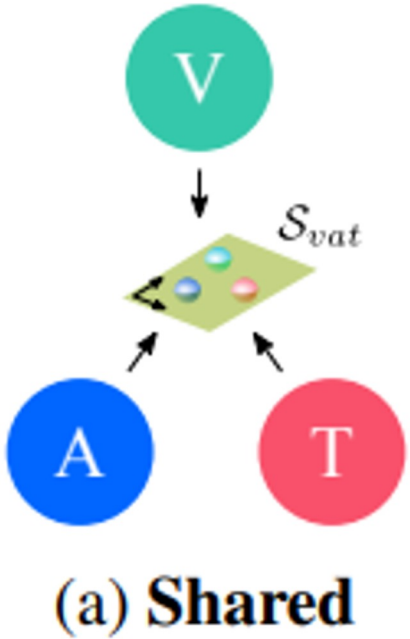
$z_{m,s} = g_{m \rightarrow s}(f_m(x_m))$

Vector representing input instance of modality m in the shared embedding space

MMV: MultiModal Versatile Networks

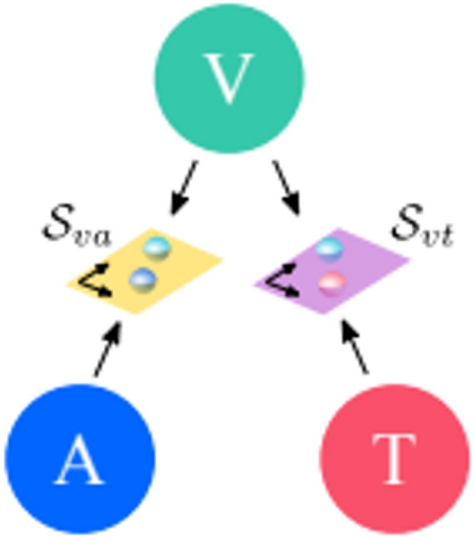
Three options considered for the modality embedding graphs:

- **Option 1: Shared space.** All modalities are embedded into a single shared vector space. Therefore only a single projection head is applied to obtain the embedding used to compare to the audio and text modalities.
    - It is easy, simple, but assumes all modalities have equal granularity (breaks property 2).



MMV: Multimodal Versatile Networks

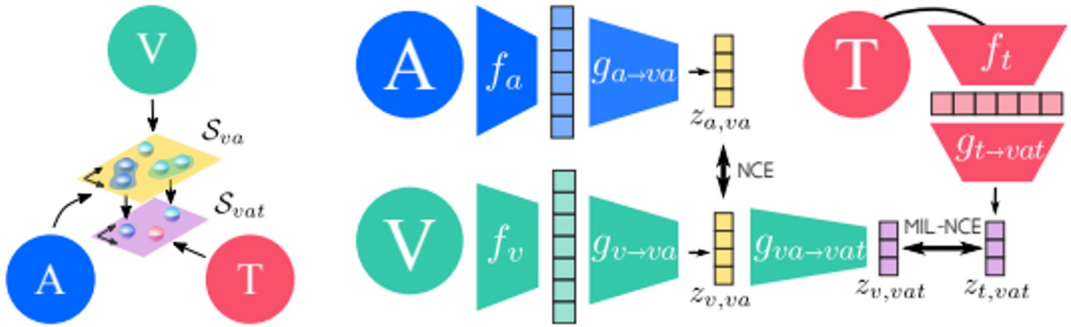
- **Option 2: Disjoint spaces.**
  Different visual-audio and visual-text spaces. There will be two distinct projection heads mappings, one for each space. However it does not follow property 3 as navigation between embedding spaces is not easily allowed.



(b) **Disjoint**

MMV: MultiModal Versatile Networks

- **Option 3: Fine and coarse spaces (FAC).** Visual and audio domains have different granularities compared to the language domain. Different spaces for each combination of modularity but define a fine-to-coarse projection $g_{v \to va}$ to embed between spaces (here to compare vision to audio). The text and audio are embedded differently (thus not breaking property 2) and easy navigation across modalities is enabled (holding property 3).



(c) **FAC** (d) **FAC details**

Multimodal Contrastive Loss

- We need to train the backbone NNs and projection heads.
  - We do not want to use manual annotations.
- Use a self-supervised tasks which aim to align pairs of modalities: vision-audio or vision-text (using a pretrained ASR for audio-text).
  - **Positive pairs** created by sampling two streams from the same location of a video.
  - **Negative pairs** created by sampling streams from different videos.
- Given these pairs, use contrastive loss to make positive pairs similar and negative pairs dissimilar in their corresponding joint embedding space.

Multimodal Contrastive Loss

- Therefore the authors minimize multimodal contrastive loss:

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

- With the component corresponding to the visual-audio pair is the following NCE loss:

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

- Correspondence between narrations and the visual channel of the video is much weaker than necessary so MIL-NCE from Miech et al. 2020 is used to account for misalignment issues.

Video to Image Network Deflation

- To comply with property 4 the authors introduce a **network deflation operation** to transform a video network into a network that can ingest a single image.
- Deflation is done by summing the 3D spatio-temporal filters over the temporal dimension to obtain 2D filets in 3D convolutional networks. In TSM networks we just need to turn off the channel shifting.
- Because these operations are zero-padded we do not achieve the desired equivalence; therefore new parameters are trained to minimize a L1 loss between output of the original video network and the output of the deflated network for the same image.

Experiments: Modality Graph Design Exploration

(a) **Benefits of multiple modalities on HT**

Modalities	UCF	HMDB	YC2	MSRVTT	ESC-50
VT	82.7	55.9	**33.6**	27.5	/
VA	75.5	51.6	/	/	**79.0**
VAT (FAC)	**84.7**	**57.3**	32.2	**28.6**	78.7
	👍	👍	?	👍	👍

HT: HowTo100M

- Learning with all three modalities outperforms with only pairs of modalities.
- On visual task (UCF101, HMDB), MMV-VAT obtains the best visual representation
- On audio task (ESC-50), MMV-VAT obtains on-par audio representation
- On visual-text task (YC2, MSRVTT), MMV-VAT wins on MSRVTT but loses on YC2

Experiments: Modality Graph Design Exploration

(b) **VAT: modality merging strategies on HT+AS**

Strategy	UCF	HMDB	YC2	MSRVTT	ESC-50
Shared	84.7	60.2	20.8	22.4	**88.5**
Disjoint	85.1	59.3	**25.0**	22.5	87.0
FAC	**86.2**	**62.5**	23.8	**23.5**	88.0
	👍	👍	👎	👎	👍

HT: HowTo100M | **AS:** AudioSet

- FAC dominates the results
- Comparing with HT pretraining only, combining AS improves performance on UCF/HMDB/ESC-50, but decrease performance on MSRVTT / YC2.
- Such decrease can be explained by the fact that only half of the training samples contain text compared to Table a

Experiments: Downstream Action Classification

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 [51]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [51]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [43]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [43]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [34]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [70]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [6]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [6]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [67]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [57]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	**89.2**		
GDT [64]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [64]	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 [21, 42, 67, 74, 90]						96.8	71.5	75.9	86.5†	43.9	81.8

Experiments: Downstream Action Classification

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 [51]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [51]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [43]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [43]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [34]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [70]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [6]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [6]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [67]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [57]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	**89.2**		
GDT [64]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [64]	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 [21, 42, 67, 74, 90]						96.8	71.5	75.9	86.5†	43.9	81.8

Experiments: Downstream Action Classification

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 [51]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [51]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [43]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [43]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [34]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [70]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [6]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [6]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [67]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [57]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	**89.2**		
GDT [64]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [64]	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 [21, 42, 67, 74, 90]						96.8	71.5	75.9	86.5†	43.9	81.8

Experiments: Downstream Action Classification

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 [51]	I3D (12.1M)	HT	15	VT	83.4	89.1	54.8	59.2	/	/	
MIL-NCE [51]	S3D-G (9.1M)	HT	15	VT	82.7	91.3	53.1	61.0	/	/	
AVTS [43]	MC3 (11.7M)	AS	1	VA		89.0		61.6	80.6		
AVTS [43]	MC3 (11.7M)	SNet	1	VA					82.3		
AA+AV CC [34]	RN-50 (23.5M)	AS	1	VA						28.5	
CVRL [70]	R3D50 (33.3M)	K600	0.1	V							64.1
XDC [6]	R(2+1)D-18 (33.3M)	AS	1	VA		91.2		61.0	84.8		
XDC [6]	R(2+1)D-18 (33.3M)	IG65M	21	VA		94.2		67.4			
ELo [67]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [57]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7	**89.2**		
GDT [64]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1	88.5		
GDT [64]	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 [21, 42, 67, 74, 90]						96.8	71.5	75.9	86.5†	43.9	81.8

Experiments: Transfer Learning to Image Classification

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 [15] ResNet50	/	ImageNet	80.5	69.3	89.0
SimCLR [15] ResNet50x2	/	ImageNet	/	74.2	92.0
SimCLR [15] ResNet50x4	/	ImageNet	84.2	76.5	93.2

def: deflation (proposed) | **n-def:** naive deflation (parameter-free)
i-inf: input inflation (repeat and stack static images)

Comparison: MMV v.s. VATT

MMV 😍

🤩 General scheme for presentation learning from multimodal data

👁️ Insightful work on modality granularity alignment

🛠️ Fixable to combine with different modality / model / loss

👍 Better result

VATT 🤔

🤖 Extension work based on MMV at the model side

🤦 Suboptimal solution with engineering tricks (Token Drop)

💸 Huge computing resource

Comparison: MMV v.s. VATT

Effect on Token Dropping Strategy

	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

performance decrease

Linear Probing Comparison

METHOD	UCF101	HMDB51	ESC50
MIL-NCE [59]	83.4	54.8	-
AVTS [50]	-	-	82.3
XDC [2]	-	-	84.8
ELo [67]	-	64.5	-
AVID [80]	-	-	**89.2**
GDT [65]	-	-	88.5
MMV [1]	**91.8**	**67.1**	88.9
VATT-Medium + SVM	89.2	63.3	82.5
VATT-Medium + LRC	89.6	65.2	84.7
VATT-MA-Medium + LRC	84.4	63.1	81.2

Visual-Text Representation Comparison

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

Comparison: MMV v.s. VATT

MMV

number of experiments, we use the S3D-G [90] network as the video backbone, with 16 frames per video clip, a total batch size of 512 and 500K training steps (20 hours training on 16 Cloud TPUs).

VATT

(BBS: 197M), Medium-Base-Small (MBS: 264M), and Large-Base-Small (LBS: 415M). Pre-training an MBS VATT with batch size 2048 on 256 TPUs (v3) takes less than 3 days. Pre-training with batch size 512 takes less than 1 day.