BraVe: Broaden Your Views for Self-Supervised Learning

Adrià Recasens[1], Pauline Luc[1], Jean-Baptiste Alayrac[1], Luyu Wang[1], Florian Strub[1], Corentin Tallec[1], Mateusz Malinowski[1], Viorica Patraucean[1], Florent Altché[1], Michal Valko[1], Jean-Bastien Grill[1], Aäron van den Oord[1], Andrew Zisserman[1,2]

[1]DeepMind, [2]VGG, University of Oxford

Code: github.com/deepmind/brave

Motivation, Goal and Contributions

Goal

- Learn good representations by regressing a broad view of the video.

Motivation

- BraVe learns strong representations of video as the narrow view needs to predict the representation of the whole video clip (broad view).
- We use separate backbones to process both views, as they perform different tasks. This enables using different augmentations/modalities in both views.
- Flow or alternative representations of the video can provide a strong signal for learning.

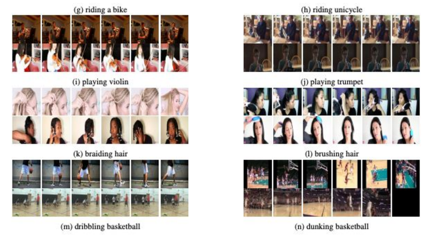
Contributions

- We propose a general framework to learn representations by predicting a broader view of the video.
- BraVe can be used with uni-modal or multi-modal data.
- Our framework enables the use of different augmentations on the different views of the video.
- We achieve SoTA results for self-supervised learning on several video and audio downstream tasks.

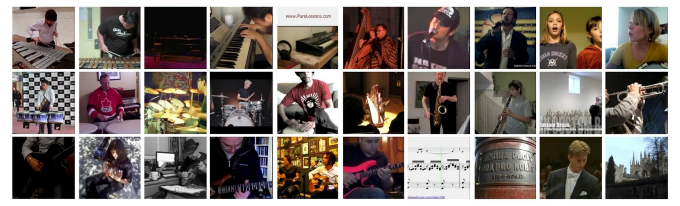
Datasets

Pre-training datasets

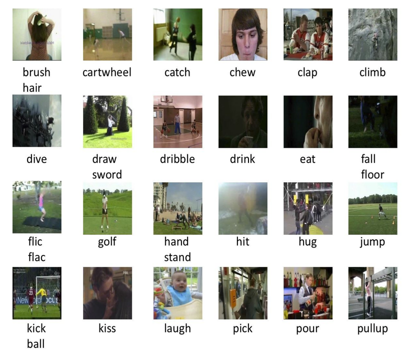
Kinetics: vision



AudioSet: vision, audio (Images from Arandjelović et al, 2018)

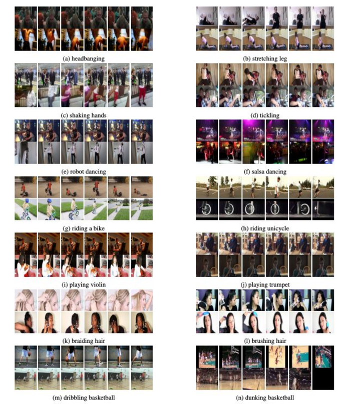


Evaluation datasets



HMDB51: 51 classes, 6.7K clips total

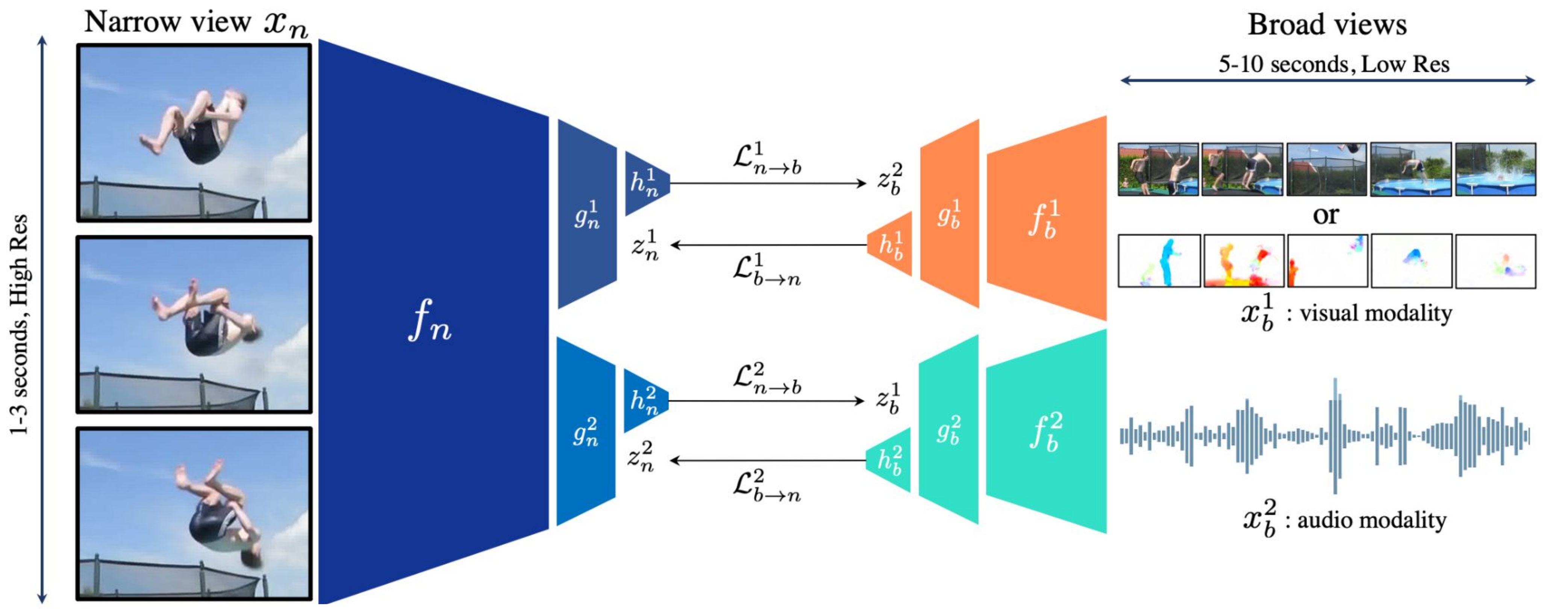
UCF101: 101 classes, 13K clips total



Kinetics-600: 600 classes, 447K clips total

Main Approach

- **BraVe architecture:** BraVe learns by regressing a representation of a broad view of the video from a narrow view.



- **BraVe loss**

$$\mathcal{L}(x) = \underbrace{\mathcal{L}_{n\to b}(x)}_{\text{Narrow}\to\text{Broad}} + \underbrace{\mathcal{L}_{b\to n}(x)}_{\text{Broad}\to\text{Narrow}}$$

$$\mathcal{L}_{b\to n}(x) = \left\| \frac{h_b(z_b)}{\|h_b(z_b)\|_2} - \mathrm{sg}\left[\frac{z_n}{\|z_n\|_2}\right] \right\|_2^2 \quad \mathcal{L}_{n\to b}(x) = \left\| \frac{h_n(z_n)}{\|h_n(z_n)\|_2} - \mathrm{sg}\left[\frac{z_b}{\|z_b\|_2}\right] \right\|_2^2$$

Research Questions

All results are linear evaluation.

Importance of the broad view

- We demonstrate that using a longer broad view on the video modality improves performance.
- When using an audio broad view, the broad view length is less relevant to final performance.

Dataset	M_b	τ_n	τ_b	HMDB51	UCF101	K600
K600	RGB+RC	10s	10s	58.7	80.0	47.4
K600	RGB+RC	1.3s	1.3s	59.4	88.1	66.3
K600	RGB+RC	1.3s	5s	61.4	88.9	65.1
K600	RGB+RC	1.3s	10s	**65.1**	**90.0**	**67.4**
AS	Audio	1.3s	1.3s	**68.3**	92.2	69.0
AS	Audio	1.3s	5s	67.5	92.4	69.9
AS	Audio	1.3s	10s	67.3	**92.6**	**70.3**

Broad view modality

- Using alternative modalities in the broad view such as randomly convoluted frames or flow improves the model performance.

M_b	HMDB51	UCF101	K600
RGB	61.3	89.9	**67.7**
RGB+RC	65.1	90.0	67.4
Flow	**65.6**	**91.1**	65.8

Syncing narrow and broad view

- Independently sampling the narrow and broad visual views results on improved performance.

Dataset	Sync	M_b	HMDB51	UCF101	K600
K600	✗	RGB+RC	**65.1**	**90.0**	**67.4**
K600	✓	RGB+RC	64.2	86.2	59.9

Number of views

- Using more than one broad view of the same modality improves the overall performance in all the benchmarks.

Dataset	Number of views	HMDB51	UCF101	K600
K600	1	65.1	90.0	67.4
K600	2	**65.6**	**91.7**	69.1
K600	3	65.2	91.5	**69.5**

Visual-only: Comparison to State-of-the-Art

- BraVe outperforms CVLR [1] when using only vision in HMDB and UCF.
- Adding flow improves performance on HMDB51 and UCF101.
- BraVe performs close to ρBYOL [2] without using EMA networks.

					UCF101		HMDB51		K600	ESC-50	AS
Method	Backbone (#params)	Dataset	Years	$\mathcal{M}$	Linear	FT	Linear	FT	Linear	Linear	MLP
CoCLR [32]	S3D (9.1M)	K400	0.07	VF	74.5	87.9	46.1	54.6		/	/
CVRL [67]	R3D50 (31.8M)	K600	0.1	V	90.6	93.4	59.7	68.0	**70.4**	/	/
ρBYOL [23]	R3D50 (31.8M)	K400	0.07	V		95.5		73.6		/	/
ρBYOL [23]	S3D (9.1M)	K400	0.07	V		**96.3**		**75.0**		/	/
BraVe:V↔V×3 (ours)	R3D50 (31.8M)	K400	0.07	V	90.6	93.7	65.1	72.0	66.5	/	/
BraVe:V↔F×3 (ours)	R3D50 (31.8M)	K400	0.07	VF	92.0	94.7	67.5	72.7	66.7	/	/
BraVe:V↔V×3 (ours)	TSM-50 (23.5M)	K600	0.1	V	91.6	94.1	65.2	73.1	69.5	/	/
BraVe:V↔F×3 (ours)	TSM-50 (23.5M)	K600	0.1	VF	91.9	94.7	65.7	74.0	67.1	/	/
BraVe:V↔V×3 (ours)	R3D50 (31.8M)	K600	0.1	V	91.9	94.4	67.6	73.9	69.1	/	/
BraVe:V↔F×3 (ours)	R3D50 (31.8M)	K600	0.1	VF	**92.7**	95.1	**68.9**	74.3	68.1	/	/

Audio-visual: Comparison to State-of-the-Art

- When using similar visual backbones and dataset, BraVe beats SOTA self-supervised models.
- When using a larger backbone, BraVe is competitive with SOTA supervised models.
- The performance of the audio backbone beats all previous self-supervised models.

					UCF101		HMDB51		K600	ESC-50	AS
Method	Backbone (#params)	Dataset	Years	$\mathcal{M}$	Linear	FT	Linear	FT	Linear	Linear	MLP
ELo [66]	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 [63]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1		88.5	
MMV [4]	R(2+1)D-18 (33.3M)	AS	1	VA	83.9	91.5	60.0	70.1	55.5	85.6	29.7
XDC [5]	R(2+1)D-18 (33.3M)	AS	1	VA		93.0		63.7		84.8	
XDC [5]	R(2+1)D-18 (33.3M)	IG65M	21	VA		95.5		68.9		85.4	
BraVe:V↔A (ours)	TSM-50 (23.5M)	AS	1	VA	**93.4**	95.6	69.1	75.3	**71.1**	92.1	**36.4**
BraVe:V↔FA (ours)	TSM-50 (23.5M)	AS	1	VFA	93.2	95.8	70.2	76.9	70.3	92.6	36.3
BraVe:V↔FA (ours)	TSM-50x2 (93.9M)	AS	1	VFA	92.8	**96.5**	**70.6**	**79.3**	70.5	**92.9**	**36.4**
Supervised [12, 44, 66, 85]						96.8	71.5	75.9	82.4	94.7	43.9

References

[1] Qian et al, Spatiotemporal contrastive video representation learning. CVPR 2021
[2] Feichtenhofer et al, *A large-scale study on unsupervised spatiotemporal representation learning.*, CVPR 2021