Scaling Multimodal Pre-Training via Cross-Modality Gradient Harmonization

Junru Wu, Yi Liang, Feng Han, Hassan Akbari, Zhangyang Wang, Cong Yu

Contrastive Multimodal Pre-Training

- Contrastive multimodal pretraining[1][2][3] on noisy multimodal video consist of ***video-audio-text*** triplets.

[1] End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020
[2] Self-Supervised MultiModal Versatile Networks, NeurIPS 2020
[3] VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. NeurIPS 2021

Motivation: Formulation of Contrastive Multimodal Pre-Training

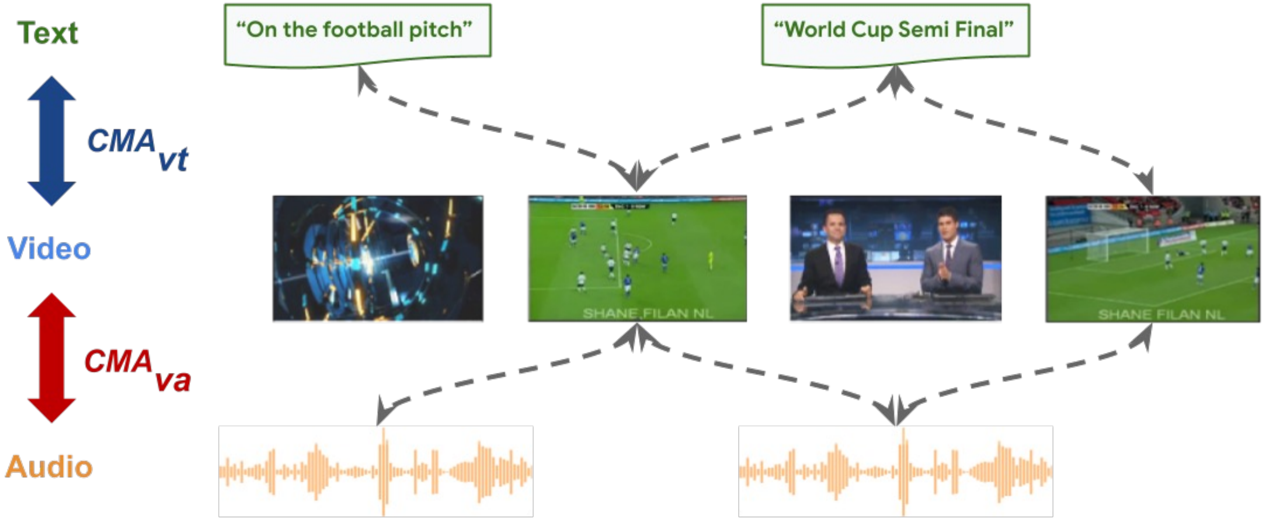
- Contrastive multimodal pretraining[1][2][3] usually consist of two pairwise contrastive losses for *video-audio* and *video-text* respectively, to solve a cross-modal alignment (CMA) problem:

$$\min_{\theta}\ \mathrm{CMA}_{va}(\theta) + \mathrm{CMA}_{vt}(\theta)$$

where *CMA_{va}* and *CMA_{vt}* penalize the cross-modal alignment for video-audio and video-text pairs, using the Noise-Contrastive Estimation objective.

$$\mathrm{CMA}_{va}(\theta) = -\log\left(\frac{\exp(z_v^\top z_a/\tau)}{\exp(z_v^\top z_a/\tau) + \sum_{z'\in\mathcal{N}} \exp({z'}_v^\top {z'}_a/\tau)}\right)$$

$$\mathrm{CMA}_{vt}(\theta) = -\log\left(\frac{\sum_{z_t\in\mathcal{P}_k(z_t)} \exp(z_v^\top z_t/\tau)}{\sum_{z_t\in\mathcal{P}_k(z_t)} \exp(z_v^\top z_t/\tau) + \sum_{z'\in\mathcal{N}} \exp({z'}_v^\top {z'}_t/\tau)}\right)$$



[1] End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020
[2] Self-Supervised MultiModal Versatile Networks, NeurIPS 2020
[3] VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. NeurIPS 2021

Motivation: Caveats in Cross-modality Alignment

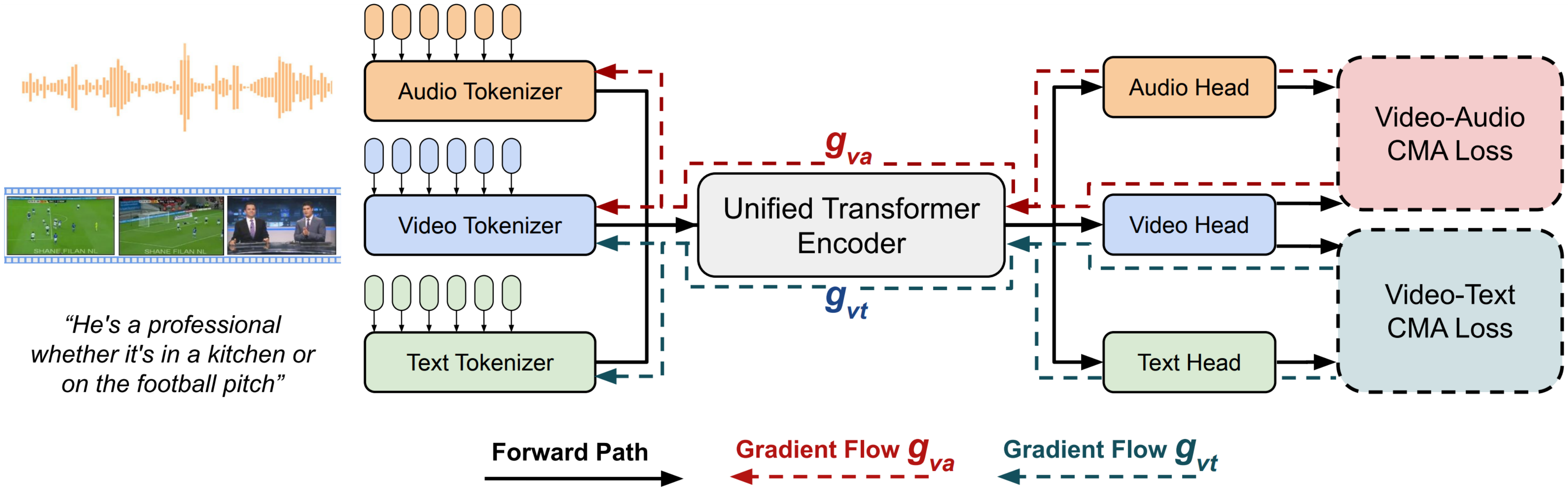
- Even on the commonly adopted instructional videos (i.e. Howto100M[1]), the cross-modality alignment (CMA) only provide weak and noisy supervision
  - e.g. a speaker can refer to something that is not visually present in the current frame, or even something irrelevant to the visual content.

Is there anyway to measure the noisiness of Cross-modality Alignment?

[1] End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

Motivation: Modality-agnostic Pre-Training

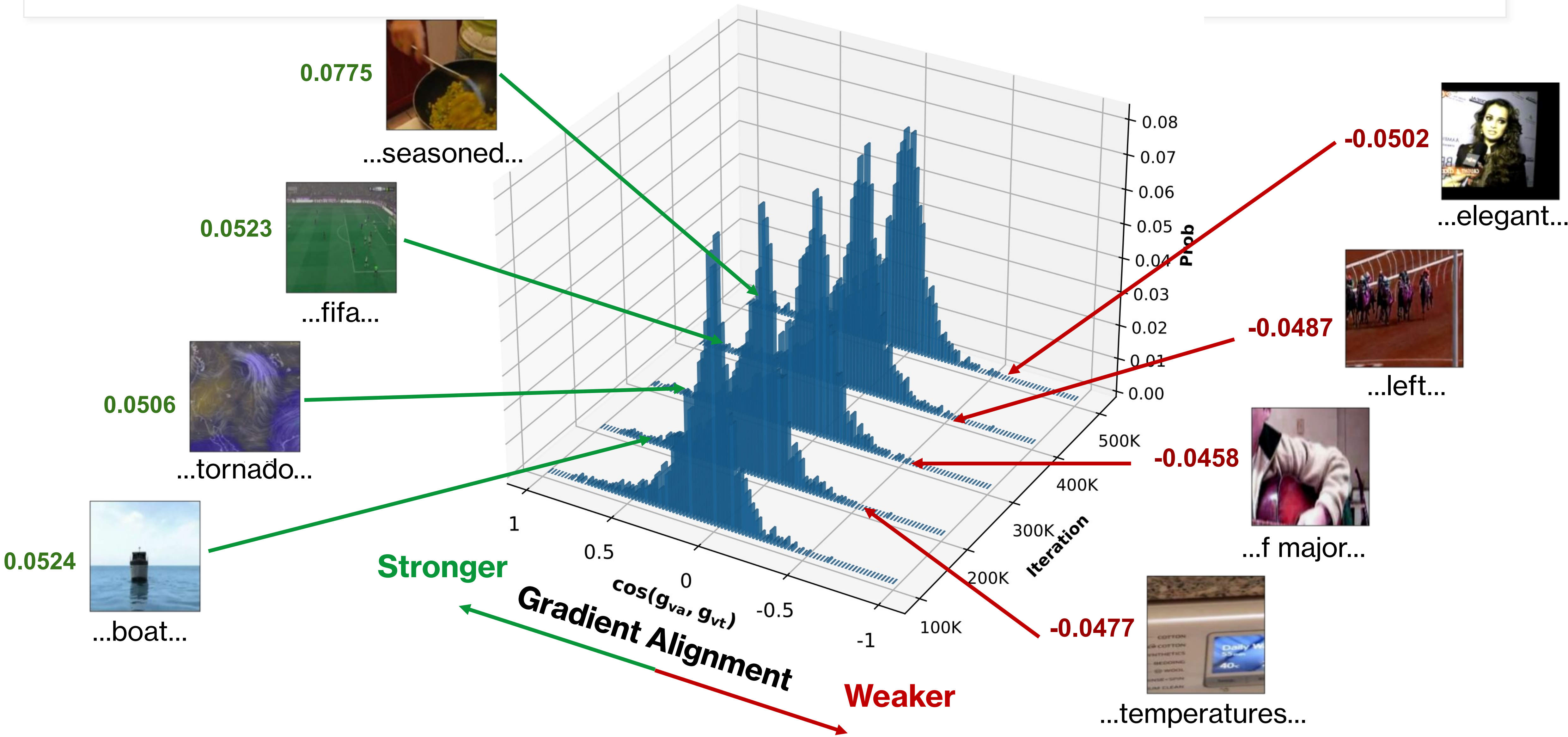
- Modality-agnostic VATT[1] as baseline
  - Gradient Conflicts[2] between g_{va} and g_{vt}
    - **Gradient Alignment** can be measured by $cos(g_{va}, g_{vt})$



Our conjecture: there is connection between **Gradient Alignment** and **Cross-Modality alignment** (CMA)

[1] VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. NeurIPS 2021
[2] Gradient surgery for multi-task learning. NeurIPS 2020

Connection between **Gradient Alignment** and **Cross-modality Alignment**?



Proposed Method: The “soft" way

Gradient-based Curriculum Learning

- We use gradient alignment *cos(g_{va}, g_{vt})* as indicator of misalignment and noisiness of samples.
- We then use **curriculum learning** to gradually identifies and removes more misaligned samples as the training goes on, based on the indicator.

Algorithm 2 Gradient-based Curriculum Learning

Require: Model parameter θ, minibatchs $\mathcal{B}_{va}$, minibatchs $\mathcal{B}_{vt}$, initial γ_0

for $(b_{va}, b_{vt}) \in (\mathcal{B}_{va}, \mathcal{B}_{vt})$ **do**

  update γ ▷ curriculumly update γ

  $g_{va} \leftarrow \nabla_\theta \text{CMA}_{va}(\theta)$, $g_{vt} \leftarrow \nabla_\theta \text{CMA}_{vt}(\theta)$

  $g_{va} \leftarrow \text{flatten}(g_{va})$, $g_{vt} \leftarrow \text{flatten}(g_{vt})$

  if $g_{va} \cdot g_{vt} > \gamma$ **then**

    $g_{va} \leftarrow \text{reshape}(g_{va})$, $g_{vt} \leftarrow \text{reshape}(g_{vt})$

    $\Delta\theta \leftarrow g_{va} + g_{vt}$ ▷ sum up gradients

    update θ with $\Delta\theta$ ▷ update parameter

  end if

end for

Proposed Method: The “hard" way

Cross-modality Gradient Realignment

- Re-align the cross-modality gradients, by re-projecting to the orthogonal direction to each other.
    - Similar to Gradient Surgery[1] originally introduced in multi-task learning.

Algorithm 1 Cross-Modality Gradient Realignment

Require: Model parameter θ , minibatchs $\mathcal{B}_{va}$, minibatchs $\mathcal{B}_{vt}$
  for $(b_{va}, b_{vt}) \in (\mathcal{B}_{va}, \mathcal{B}_{vt})$ **do**
    $g_{va} \leftarrow \nabla_\theta \text{CMA}_{va}(\theta)$, $g_{vt} \leftarrow \nabla_\theta \text{CMA}_{vt}(\theta)$
    $g_{va} \leftarrow \text{flatten}(g_{va})$, $g_{vt} \leftarrow \text{flatten}(g_{vt})$
    $\hat{g}_{va} \leftarrow g_{va}$, $\hat{g}_{vt} \leftarrow g_{vt}$
    if $g_{va} \cdot g_{vt} < 0$ **then**
      $\hat{g}_{va} \leftarrow \hat{g}_{va} - \frac{\hat{g}_{va} \cdot g_{vt}}{||g_{vt}||^2}$ ▷ projection $g_{vt} \leftarrow g_{va}$
      $\hat{g}_{vt} \leftarrow \hat{g}_{vt} - \frac{\hat{g}_{vt} \cdot g_{va}}{||g_{va}||^2}$ ▷ projection $g_{va} \leftarrow g_{vt}$
    end if
    $\hat{g}_{va} \leftarrow \text{reshape}(\hat{g}_{va})$, $\hat{g}_{vt} \leftarrow \text{reshape}(\hat{g}_{vt})$
    $\Delta\theta \leftarrow \hat{g}_{vt} + \hat{g}_{va}$ ▷ sum up gradients
    update θ with $\Delta\theta$ ▷ update parameter
  end for

[1] Gradient surgery for multi-task learning. NeurIPS 2020

Experiments

GR: Gradient Realignment

CL: Gradient-based Curriculum Learning

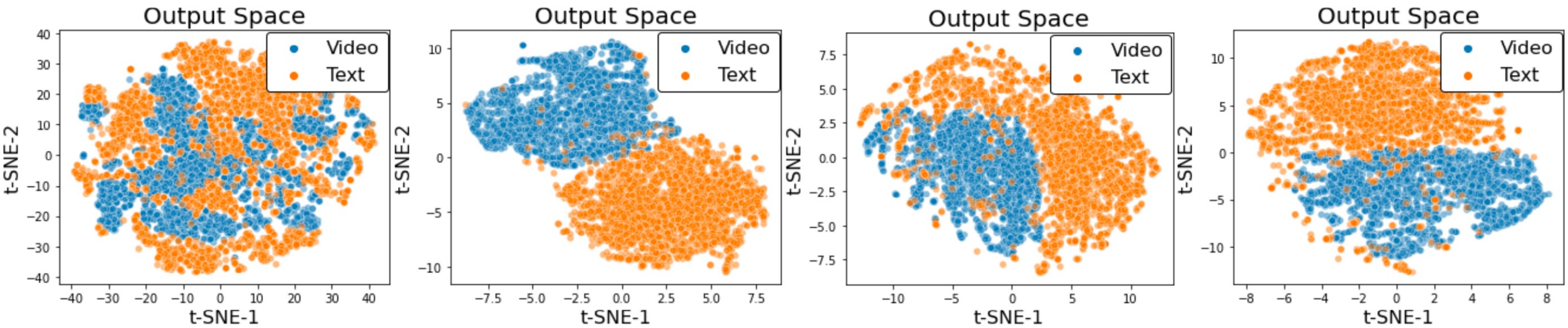
- Pre-Training
  - Howto100M
  - AudioSet
  - Youtube8M
- Downstream
  - Uni-Modal
    - Video Cls
      - UCF101
      - HMDB51
      - Kinetics400
    - Audio Cls
      - ECS50
  - Cross-Modal
    - Text-Video Retrieval
      - YouCook2
      - MSRVTT

Dataset	Tasks	Video Action Cls						Text-Video Retrieval				Audio Cls	
	Dataset	UCF101		HMDB51		Kinetics400		YouCook2		MSRVTT		ESC50	
	Metric	Top1↑	Top5↑	Top1↑	Top5↑	Top1↑	Top5↑	Rank↓	R@10↑	Rank↓	R@10↑	Top1↑	Top5↑
HT100M	VATT [11]	78.53	95.24	57.36	86.07	74.71	92.69	93.50	17.10	73.00	16.70	71.50	91.75
	+ RW (VA)	78.24	95.01	**58.74**	85.39	70.55	90.41	168.90	9.24	100.20	13.56	71.56	93.02
	+ RW (VT)	78.03	95.35	58.23	**86.80**	75.66	92.80	90.35	19.26	62.50	18.02	71.29	91.38
	+ GR	78.44	95.37	54.38	83.64	**76.72**	92.72	47.00	24.94	68.00	19.20	69.00	93.00
	+ CL	**79.10**	**96.16**	56.41	86.06	**77.25**	**93.38**	**40.00**	**26.01**	**56.00**	**23.90**	**72.50**	**94.25**
	+ Both	**79.24**	**96.58**	**58.24**	**87.37**	76.59	**93.26**	**42.00**	**25.87**	**54.00**	**24.50**	**72.05**	**94.16**
HT100M + AudioSet	VATT [11]	84.40	-	63.10	-	79.23	94.30	34.00	29.00	**67.00**	**23.60**	81.20	-
	+ RW (VA)	83.44	97.28	59.55	88.02	76.56	93.52	238.50	6.80	147.00	12.30	81.75	97.25
	+ RW (VT)	84.42	97.49	62.30	86.61	78.59	94.17	76.00	18.72	120.00	15.20	80.75	96.75
	+ GR	84.77	97.38	62.30	**90.38**	**79.29**	94.32	**29.00**	**31.65**	70.00	**21.40**	81.50	**97.00**
	+ CL	**86.04**	**97.75**	**65.45**	88.94	**79.89**	**94.71**	33.00	29.17	**65.50**	20.97	**82.00**	**97.25**
	+ Both	**85.46**	**97.58**	**65.52**	**89.74**	79.26	**94.48**	**31.50**	**30.26**	69.50	19.96	**82.00**	**97.00**
HT100M + AudioSet + YT8M	VATT [11]	88.28	**98.73**	**65.84**	91.43	79.39	94.56	**29.00**	29.66	56.00	26.90	80.75	97.00
	+ RW (VA)	86.97	98.09	61.06	89.66	77.70	93.83	99.00	14.25	75.50	19.70	83.50	97.25
	+ RW (VT)	88.19	97.96	61.13	90.51	78.43	94.38	**27.00**	31.07	48.50	27.70	82.25	96.75
	+ GR	87.49	98.10	60.99	88.35	**79.73**	94.57	32.00	29.56	60.00	27.20	**85.00**	**98.00**
	+ CL	**89.02**	98.33	**65.77**	**92.15**	79.70	**94.80**	31.00	**31.34**	**48.50**	**28.70**	83.50	**97.75**
	+ Both	**89.70**	**98.35**	64.35	**92.08**	**80.01**	**94.69**	**29.00**	**31.86**	**45.00**	**29.10**	**84.50**	**98.00**

GR + CL Improve on both uni-modal tasks and cross-modal tasks

Current SoTA on modality-agnostic setting.

Visualization: VATT output space



(a) VATT Baseline (b) w/ Gradient Realignment (c) w/ Curriculum Learning (d) w/ Both Techniques

VATT Baseline: Video and Text mixed together

w/ Our Methods: Video and Text become disentangled

Thanks for your Attention!

Project Website

Author Homepage