Turbo Training with Token Dropout

Tengda Han[1], Weidi Xie[1,2], Andrew Zisserman[1]

[1]Visual Geometry Group, University of Oxford

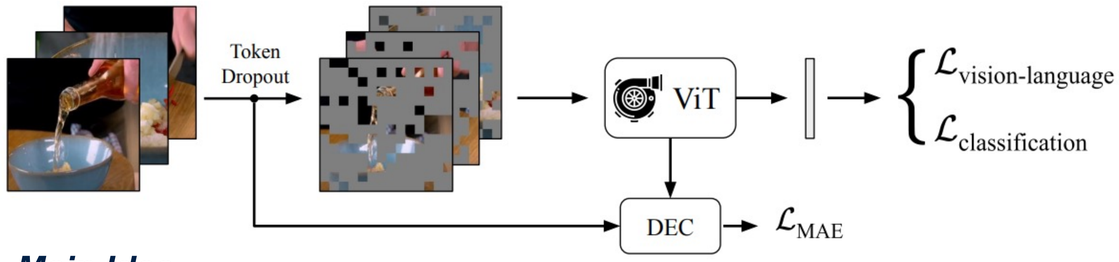
[2]Shanghai Jiao Tong University

Contributions

(i) Propose **Turbo Training**, a simple and versatile training paradigm for Transformers on multiple video tasks.

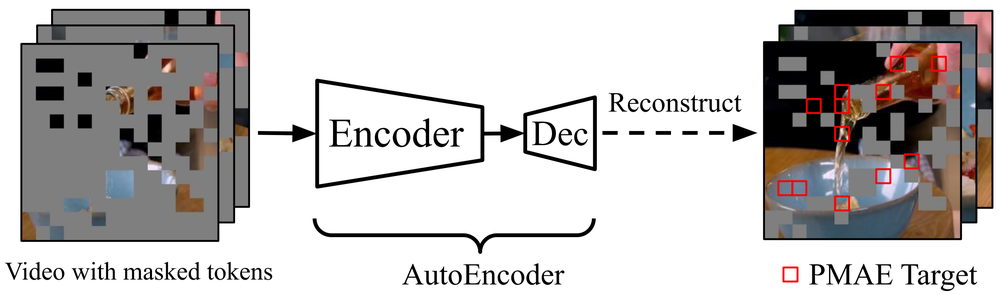
(ii) On action classification, video-language representation learning, and long-video activity classification, Turbo training achieves almost 4× speed-up and significantly less memory consumption, while largely maintaining competitive performance.

(iii) Turbo training enables long-schedule video-language training and end-to-end long-video training, delivering competitive or superior performance than previous works, which were infeasible to train under limited resources.



Main Idea

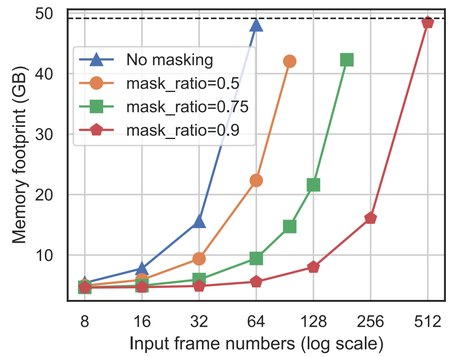
- Randomly drop visual tokens when training Vision Transformers (ViT), to save computation
- Use Masked Autoencoder as the auxiliary loss when training other tasks



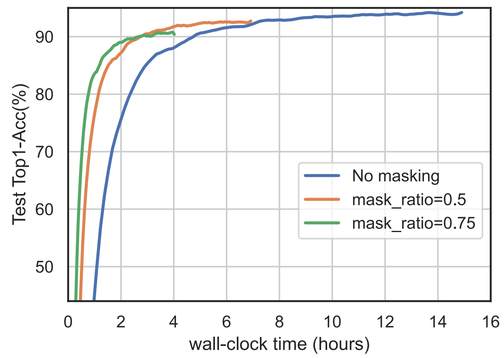
Partial-MAE

- Classic MAE is trained to reconstruct **all** the masked visual patches, which is unnecessary under a high masking ratio
- We propose Partial-MAE, that only reconstructs part of the masked patches, further save the computation

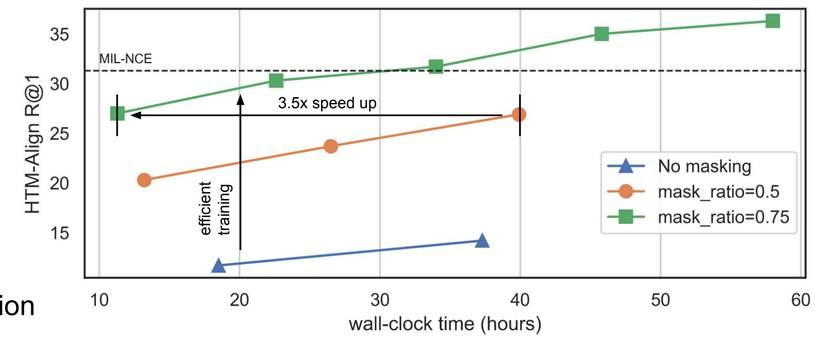
Experiments

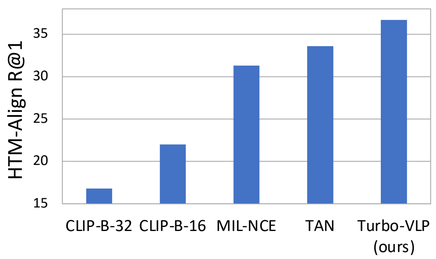


Memory-efficient: Turbo Training largely reduce the memory cost for video training, therefore enables larger batch size and more input frames



Fast Finetuning: Turbo Training speeds up the action classification finetuning by 4x, whereas still largely maintains the competitive performance.





Text-visual alignment on HTM-Align

Efficient visual-language pretraining (VLP): Turbo Training enables visual-language pretraining on large dataset (HowTo100M) for long schedule with limited resource. To reach the same downstream performance, training with 0.75 mask ratio gives 3.5x speed up compared with 0.5 mask ratio, and is much more efficient than training without token dropout.

	End-to-end	Video clip features	Breakfast	COIN
Timeception	❌	3D-ResNet	71.3	-
GHRM	❌	I3D	75.5	-
Dist. Sup.	❌	TimeSformer	89.9	**88.9**
Turbo Training	✅	N/A (end-to-end)	**91.3**	87.5

Enable end-to-end long-video task: Turbo Training supports training with more video frames within the same memory. Therefore, it enables end-to-end training on long-video action classification task – which were trained in two stages on the pre-extracted video clip features. We achieve state-of-the-art or competitive performance on Breakfast and COIN datasets.