VATT : Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Authors : Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, Boqing Gong

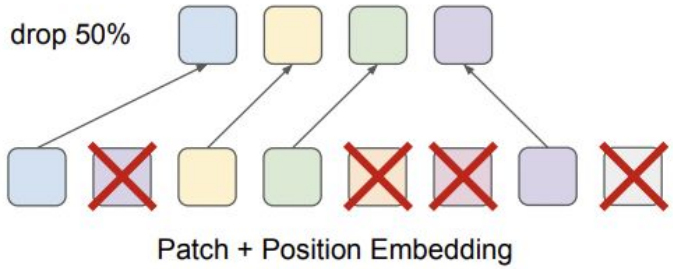
Submitted Apr 22, 2021; Last Revised Dec 7, 2021 (NeurIPS 2021)

Presented by : David K., Li Hui, Junjie

Arguments

Argument #1 - General Purpose Architecture

- A single, versatile and general purpose architecture for ALL modalities
- Minimal changes to the vision transformer so that the learned model can transfer its weights to various frameworks and tasks
- DropToken reduces FLOPs quadratically → low complexity



DropToken

Argument #2 - Self-supervised Learning (SSL) from video dataset is important

- Video datasets are relatively small compared with image datasets
- SSL let video models be **trained from scratch on large-scale, unlabeled data**
- Previous method (TimeSFormer) need to be first pre-trained on ImageNet, the model are **naturally biased by image-based models**

Argument #2 - Self-supervised Learning (SSL) from video dataset is important

- VATT is the **first** ViT backbone that is **pre-trained from scratch** using self-supervision on multimodal videos and achieves state-of-the-art results on video action recognition

framework	model	params	acc. (%)
MoCo v3	ViT-B	86M	76.7
MoCo v3	ViT-L	304M	77.6
MoCo v3	ViT-H	632M	78.1
MoCo v3	ViT-BN-H	632M	79.1
MoCo v3	ViT-BN-L/7	304M	**81.0**

MoCo v3, submitted Apr 2021 (DINO submitted Apr 2021)

Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

Argument #3 - VATT shows more significant performance improvements

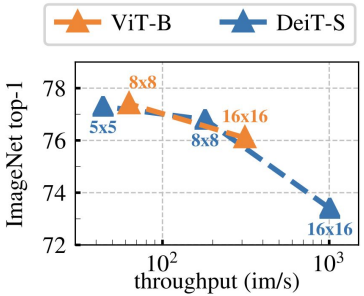
Method	Kinetics-400 Top-1	Kinetics-400 Top-5	Kinetics-600 Top-1	Kinetics-600 Top-5
I3D [13]	71.1	89.3	71.9	90.1
R(2+1)D [26]	72.0	90.0	-	-
bLVNet [27]	73.5	91.2	-	-
S3D-G [96]	74.7	93.4	-	-
Oct-I3D+NL [20]	75.7	-	76.0	-
D3D [83]	75.9	-	77.9	-
I3D+NL [93]	77.7	93.3	-	-
ip-CSN-152 [87]	77.8	92.8	-	-
AttentionNAS [92]	-	-	79.8	94.4
AssembleNet-101 [77]	-	-	-	-
MoViNet-A5 [47]	78.2	-	82.7	-
LGD-3D-101 [69]	79.4	94.4	81.5	95.6
SlowFast-R101-NL [30]	79.8	93.9	81.8	95.1
X3D-XL [29]	79.1	93.9	81.9	95.5
X3D-XXL [29]	80.4	94.6	-	-
TimeSFormer-L [9]	80.7	94.7	82.2	95.6
VATT-Base	79.6	94.9	80.5	95.5
VATT-Medium	81.1	**95.6**	82.4	96.1
VATT-Large	**82.1**	95.5	**83.6**	**96.6**
VATT-MA-Medium	79.9	94.9	80.8	95.5

VATT on video recognition accuracy

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**
Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

DINO on image classification accuracy

Throughput drops significantly when switching to 8 by 8



Argument #4 - Scalability

- VATT shows that performance can be improved when scaling up the model
- DINO only trained with relatively small backbones: ViT-S and ViT-B

Model	Layers	Hidden Size	MLP Size	Heads	Params
Small	6	512	2048	8	20.9 M
Base	12	768	3072	12	87.9 M
Medium	12	1024	4096	16	155.0 M
Large	24	1024	4096	16	306.1 M

ViT settings tested in VATT experiments

Paper Battle:
Arguments in favour of DINO

Argument 1: Motivation

- DINO has been proposed with the motivation of interpreting self-supervised ViTs, that bring forth very unique properties (scene layout and powerful features), something not done before.
- Meanwhile VATT is more of an application paper that simply adopts self-supervised learning to multimodal data, without any motivation to study insights in the paradigm.
- Thus, DINO probably makes a bigger contribution to the scientific community!

Argument 2: Quality of features learned

- DINO achiever higher top-1 accuracy on ImageNet compared to VATT
- Even the simple k-NN evaluation on raw features is very close (1%) to that obtained by VATT after pre-training + fine-tuning.
- Thus, we can say that DINO yields richer features compared to VATT

Method	Arch.	Param.	im/s	Linear	k-NN
DINO	ViT-B/8	85	63	**80.1**	77.4

Table 2: **Linear and k-NN classification on ImageNet.**

METHOD	PRE-TRAINING DATA	TOP-1	TOP-5
VATT-Base	HowTo100M	78.7	93.9

Table 3: Finetuning results for ImageNet classification.

Argument 3: Negative-free self-supervised approach

- DINO is non-contrastive whereas VATT is contrastive
- Hence, VATT requires larger batch sizes so that it has a lot of negative pairs, while that is not required in DINO
- Reducing batch size leads to severe performance drops for VATT (right) but not much for DINO (left)

bs	128	256	512	1024
top-1	57.9	59.1	59.6	59.9

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

Argument 4: Impact on research community

- Cited 7x more than VATT; significantly greater research impact across the vision community (blogs, discussions, podcasts, videos…!)

Emerging properties in self-supervised vision transformers

M Caron, H Touvron, I Misra, H Jégou, J Mairal, P Bojanowski, A Joulin

Proceedings of the IEEE/CVF international conference on …, 2021 • openaccess.thecvf.com

Abstract

In this paper, we question if self-supervised learning provides new properties to Vision Transformer (ViT) that stand out compared to convolutional networks (convnets). Beyond the fact that adapting self-supervised methods to this architecture works particularly well, we make the following observations: first, self-supervised ViT features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. Second, these features are also excellent

SHOW MORE

☆ Save Cite Cited by 3625 Related articles All 18 versions »

Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text

H Akbari, L Yuan, R Qian, WH Chuang, SF Chang, Y Cui, B Gong

Advances in Neural Information Processing Systems, 2021 • proceedings.neurips.cc

Abstract

We present a framework for learning multimodal representations from unlabeled data using convolution-free Transformer architectures. Specifically, our Video-Audio-Text Transformer (VATT) takes raw signals as inputs and extracts multimodal representations that are rich enough to benefit a variety of downstream tasks. We train VATT end-to-end from scratch using multimodal contrastive losses and evaluate its performance by the downstream tasks of video action recognition, audio event classification, image

SHOW MORE

☆ Save Cite Cited by 507 Related articles All 8 versions »