EsViT: Efficient Self-supervised Vision Transformers for Representation Learning

-- Unleash the power of unlabeled big visual data

June, 2021

Chunyuan Li
Deep Learning Team
Microsoft Research, Redmond

Microsoft

Acknowledgements to the V-Team Members:

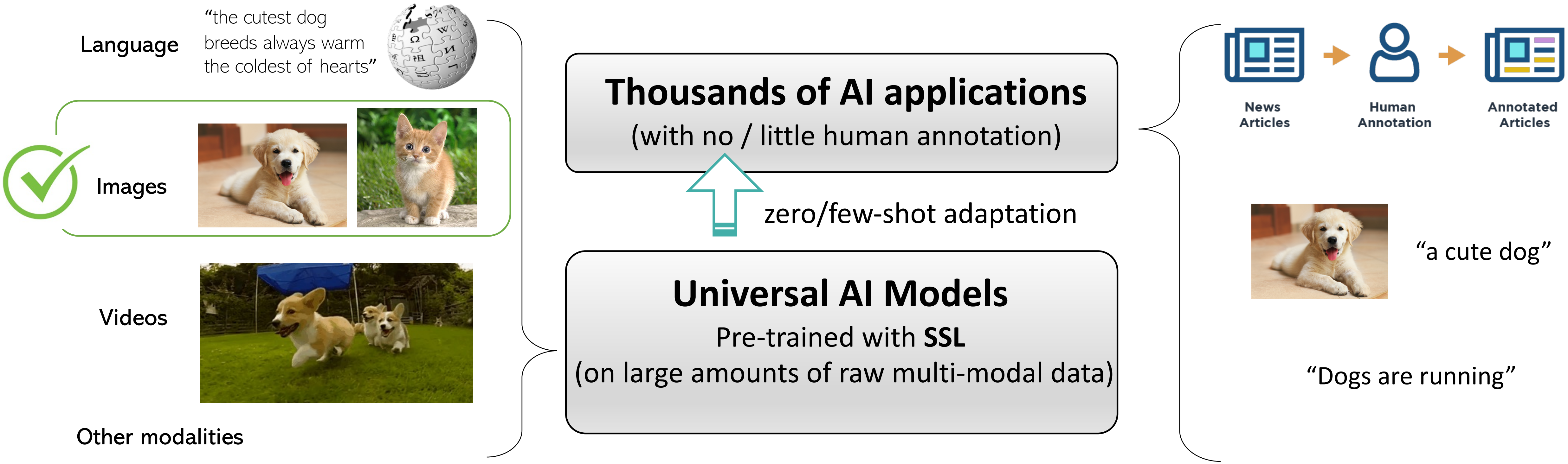
Chunyuan Li[1] **Jianwei Yang**[1] **Pengchuan Zhang**[1] **Mei Gao**[2] **Bin Xiao**[2] **Xiyang Dai**[2]
Lu Yuan[2] **Jianfeng Gao**[1]
[1]Microsoft Research at Redmond, [2]Microsoft Cloud + AI
{chunyl,jianwyan,penzhan,xuga,bixi,xidai,luyuan,jfgao}@microsoft.com

Outline of this presentation

- Project Background & Motivations
- EsViT Method
- Results
- Future Works

Why Self-Supervised Learning (SSL)?

- EsViT is a part of the bigger picture on Universal Multimodal Representation Learning
- Leveraging big unlabeled data to learn universal representations for a large range of downstream tasks



SSL: A scaling success path

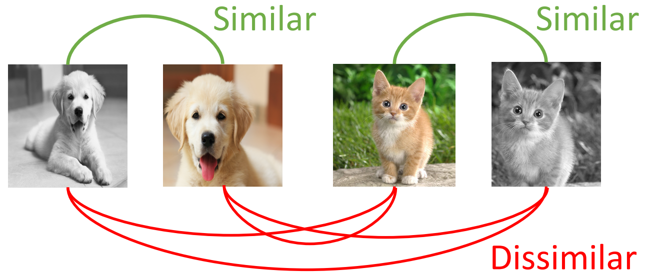
- A proved path: Scaling success in NLP
- SSL in CV ?
  - Repeat the success in NLP
  - Unleash the power of big unlabeled visual data

- Three critical ingredients in the successful recipe:
  - **Data:** Easy to collect large amounts of raw data
  - **(Pre-)Training objectives:** SSL enables the use of large no human-labeled data
  - **Network architectures:** Transformers allows efficient training of large models

Scalable

NLP
Web text corpus
Masked Token Modeling
Transformers

The current SoTA of SSL for **Images** ?

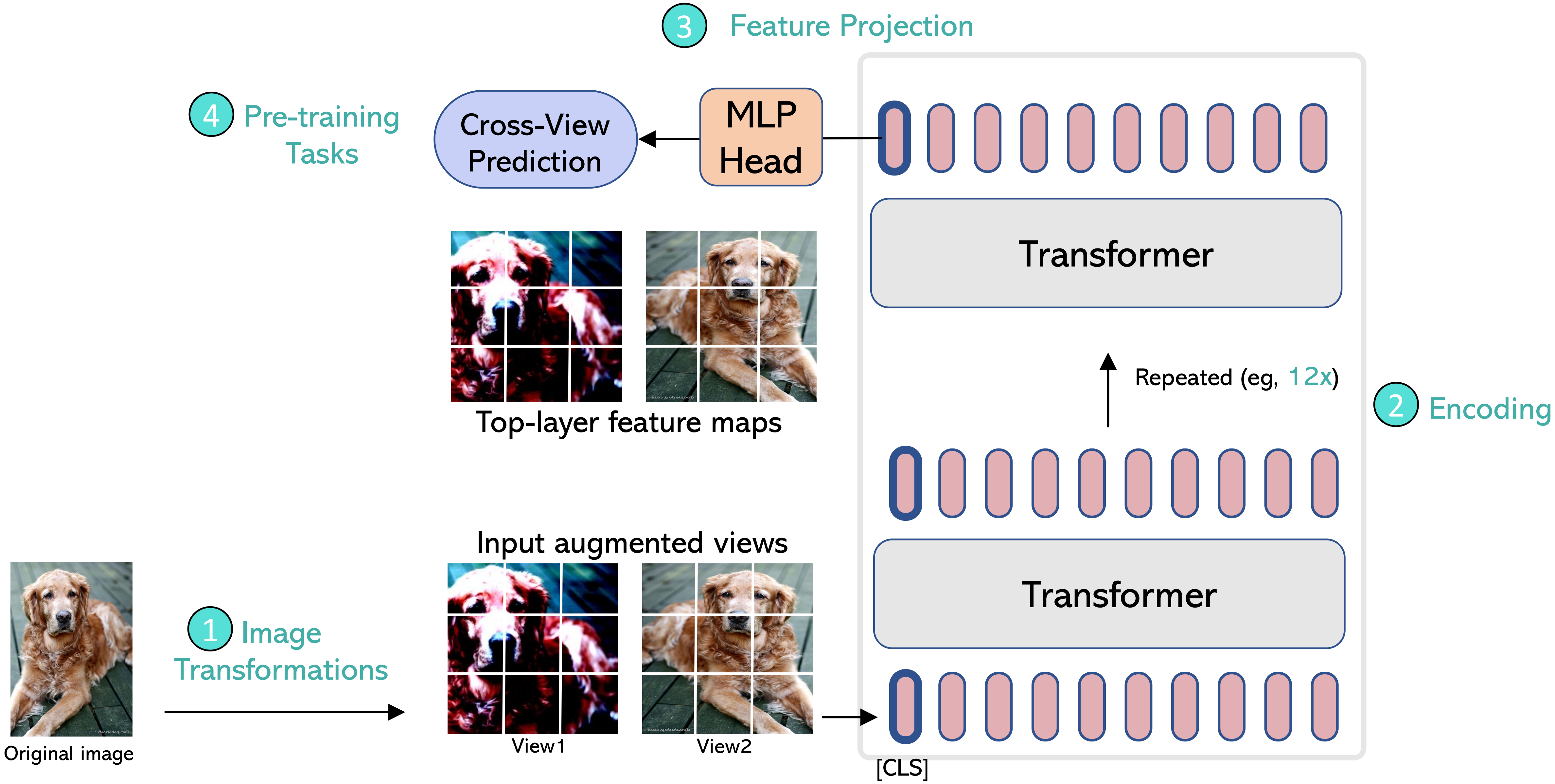
	NLP	CV
Data	• Web text corpus	• Web images
Network architectures	• Transformers • Largest model: 1.75T	• CNNs --> Transformers • Largest model: smaller than 1B
Pre-training objectives	• Sentence-level Contrastive Learning: Next Sentence Prediction • Local dependencies: Masked Token Modeling "the cutest dog breeds always ~~warm~~ the coldest of hearts"	• View-Level Contrastive Learning Augmented Views  • Local region dependencies?

The proposed method: EsViT

Network architectures: A multi-stage Transformer architecture

Pre-training Objectives: A region-level pre-train task

Monolithic Transformer Architecture (Baseline)



Multi-stage Transformer Architecture (proposed)

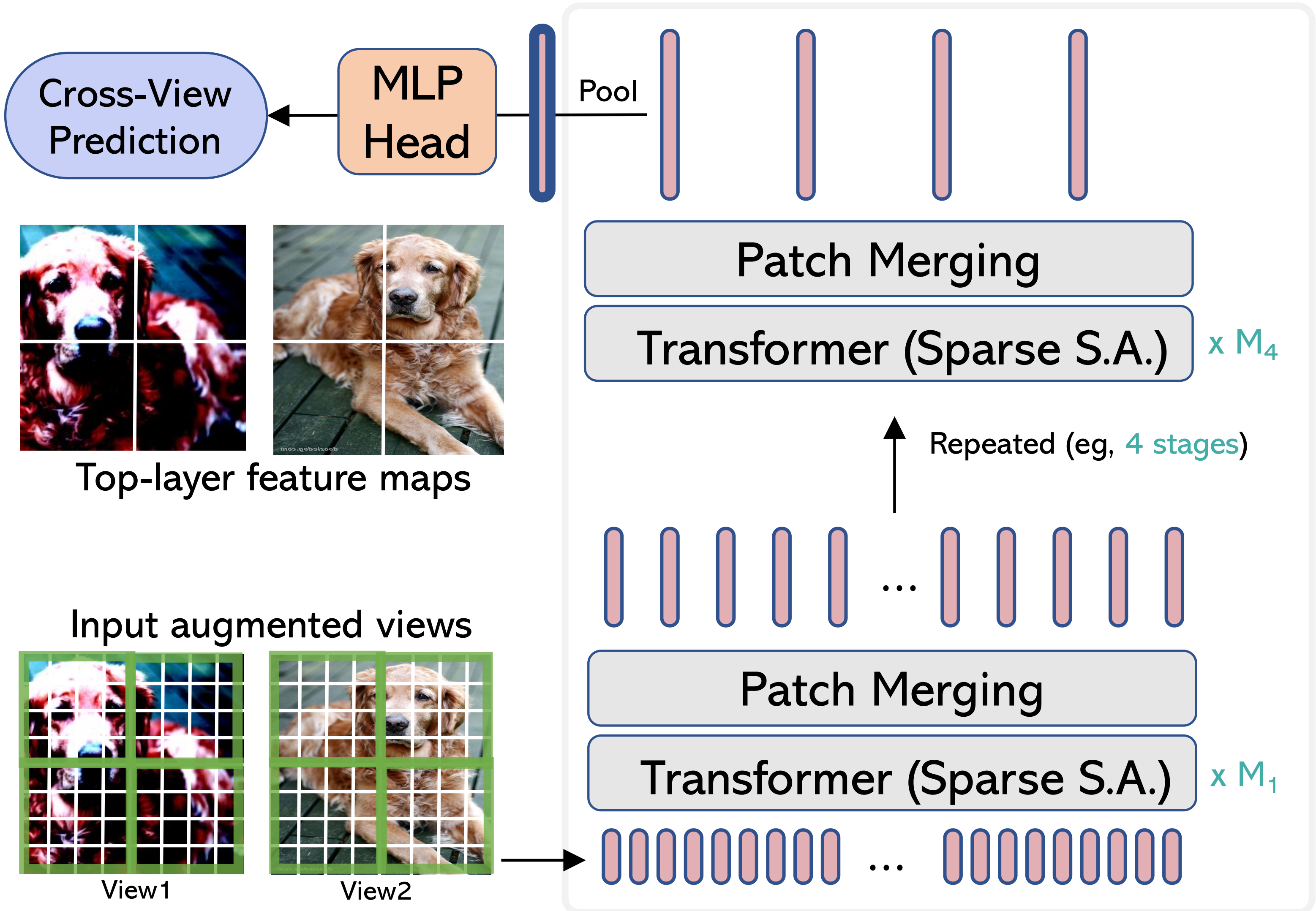
4-stage: 2-2-6-2
The number of Transformer in each stage

Reduce compute complexity !

1. Sparse Self-Attention (S.A.)
2. Merging tokens for shorter sequences

Input Views:

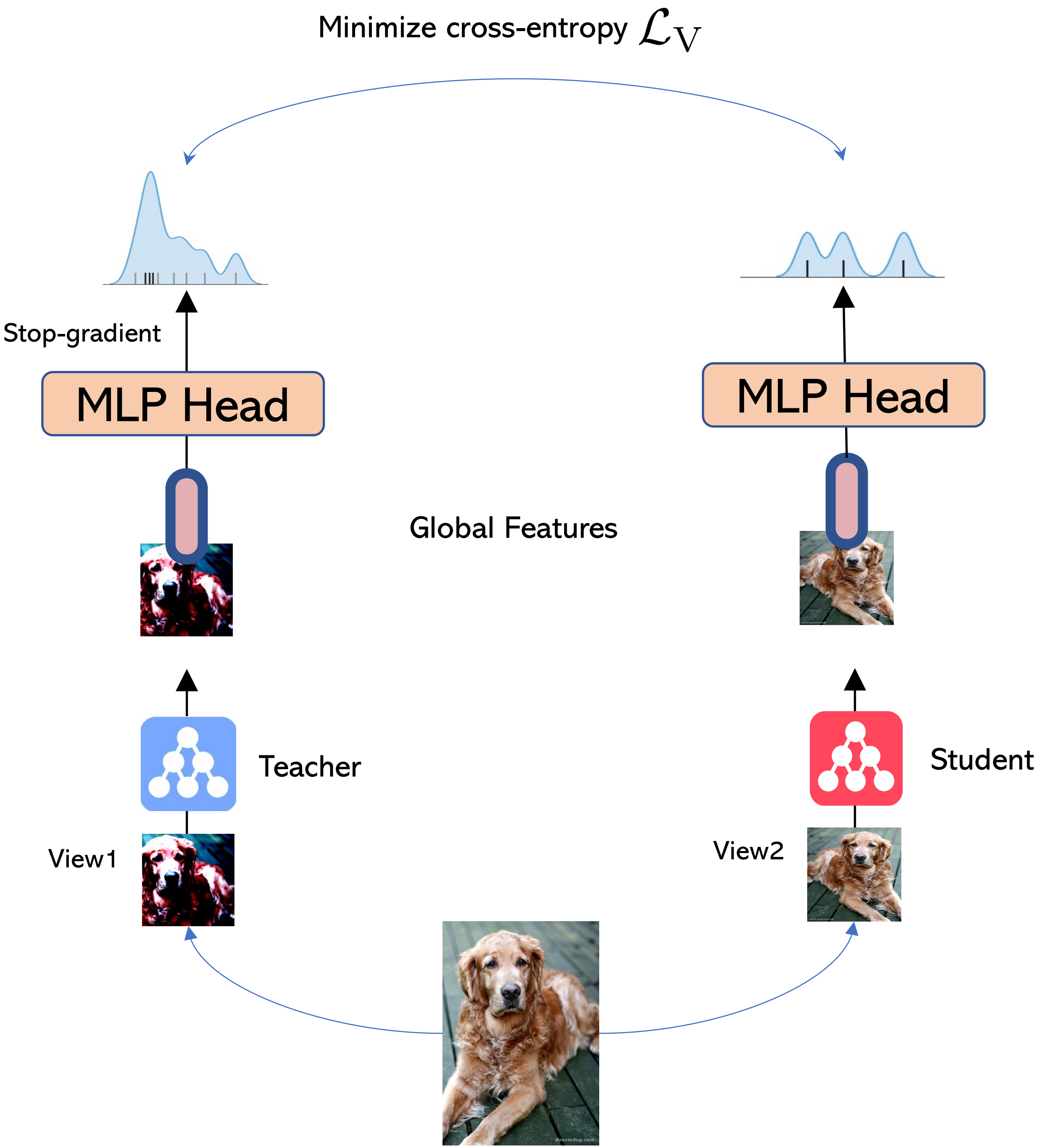
- Smaller patches & Longer sequences



Pre-train Task 1: view-level

Model updates:

- Student: SGD w.r.t. $\mathcal{L}_{\mathrm{V}}$
- Teacher: exponential moving average of student weights



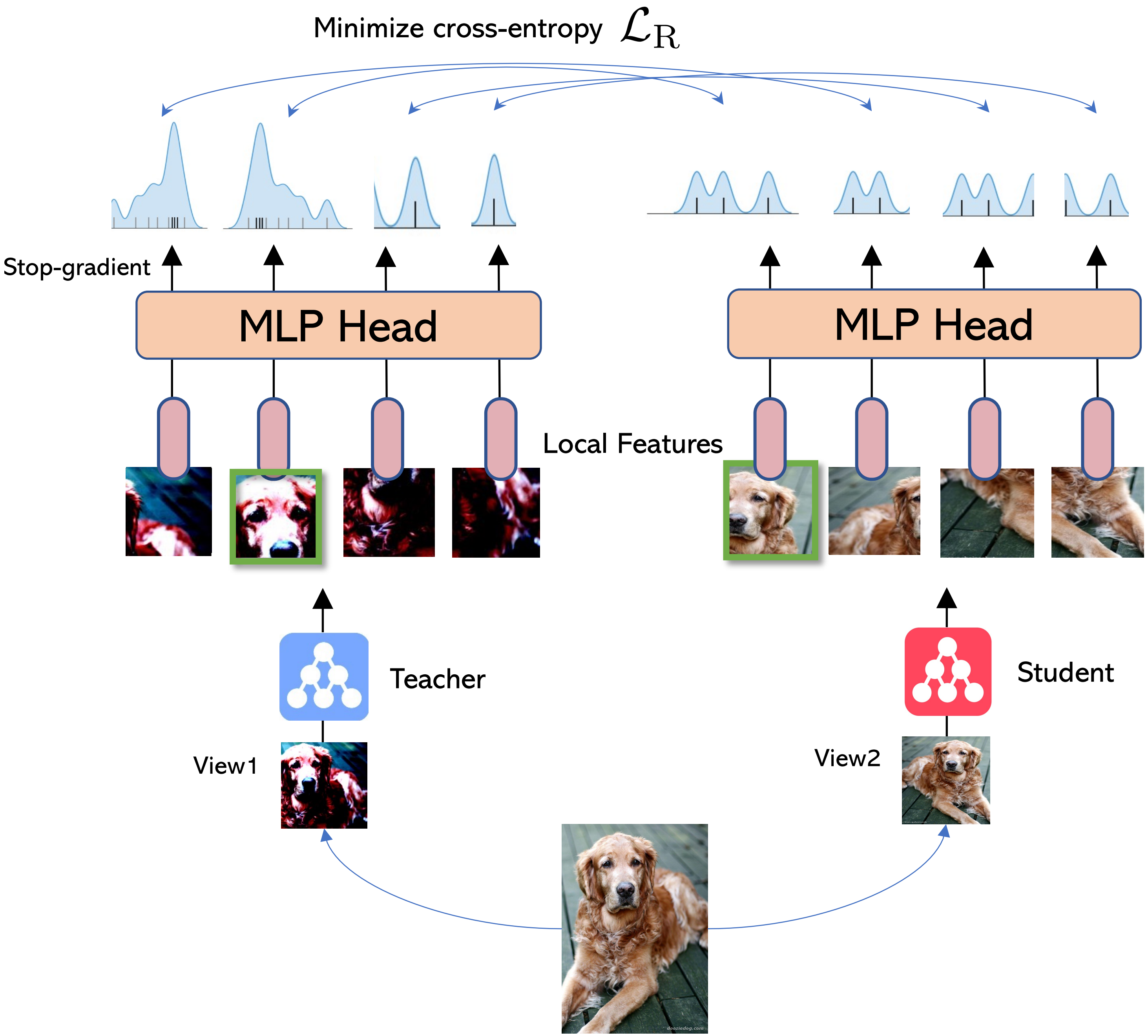
Source: Emerging Properties in Self-Supervised Vision Transformers, 2021

Pre-train Task 2: region-level

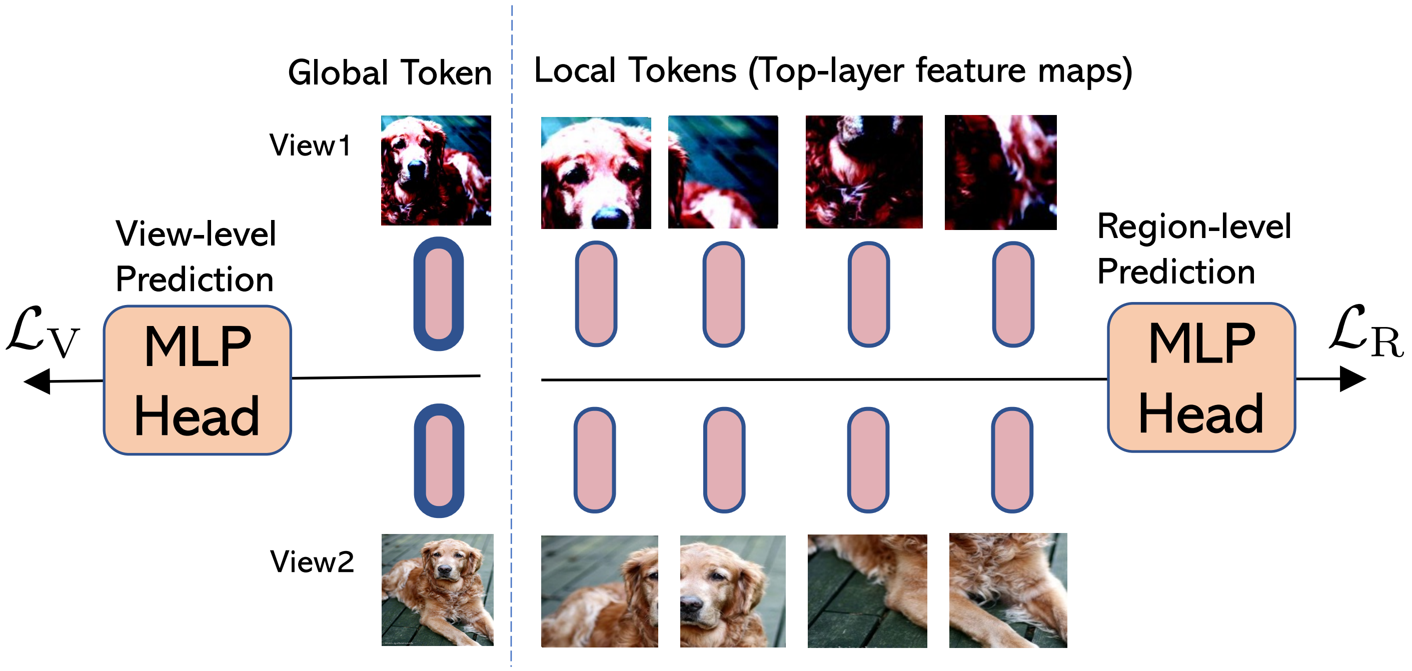
- Compute the cross-entropy between two most similar regions

- An analogy to masked token modeling in BERT:

For a region in a different augmented view, we predict its soft-label provided by the teacher model



Pre-train Tasks: both view- and region-level objectives



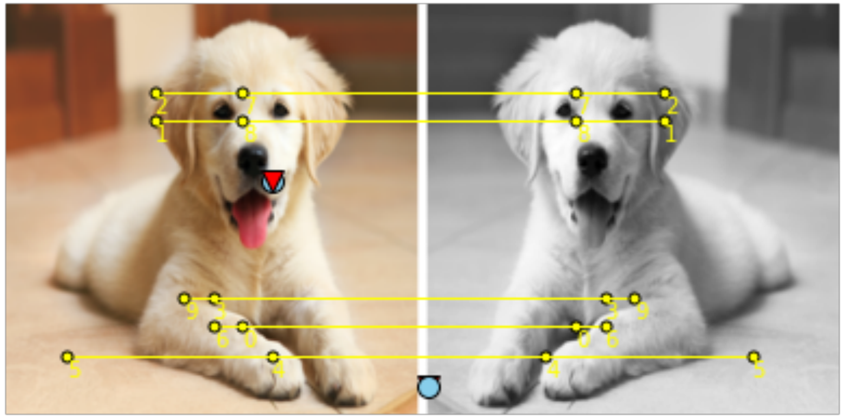
Results of EsViT

An intriguing property

Leaderboard Results

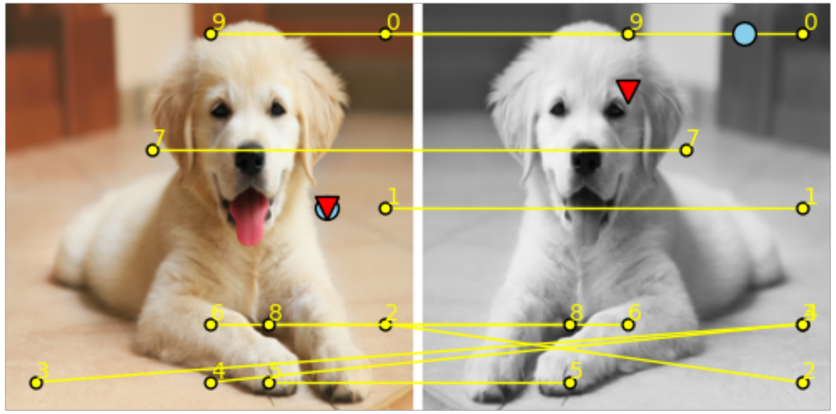
Transfer Learning

An intriguing Property of self-supervised Transformers



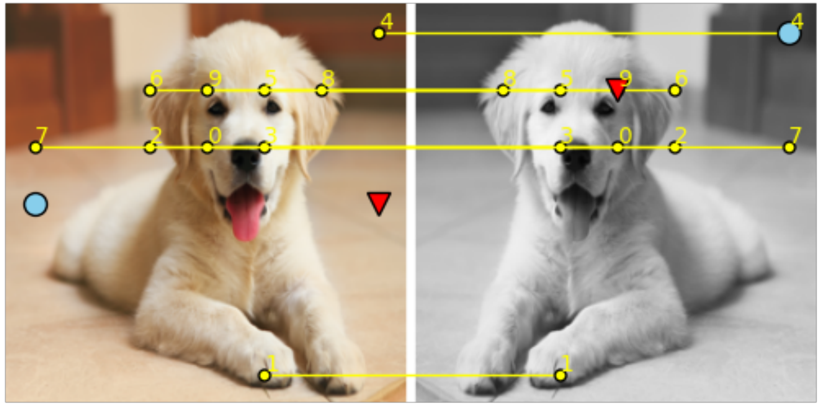
(a) DINO: Monolithic with $\mathcal{L}_{\mathrm{V}}$

Slow



(b) EsViT: Multi-stage with $\mathcal{L}_{\mathrm{V}}$

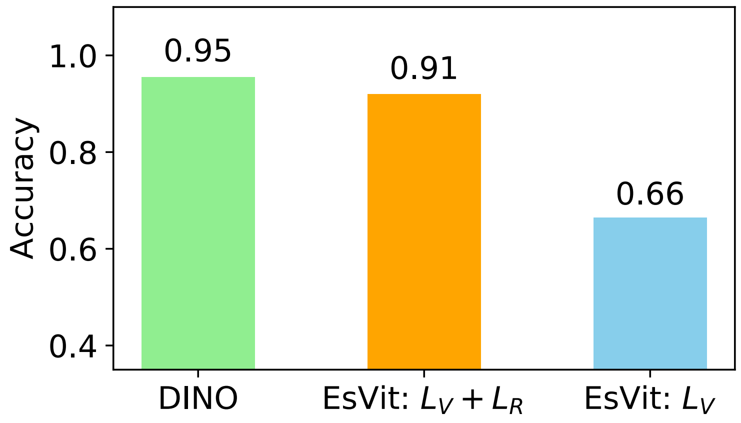
Fast

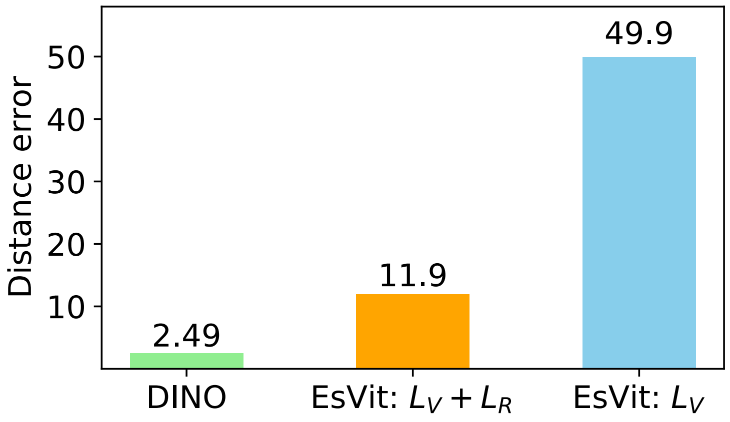


(c) EsViT: Multi-stage with $\mathcal{L}_{\mathrm{V}}$ and $\mathcal{L}_{\mathrm{R}}$

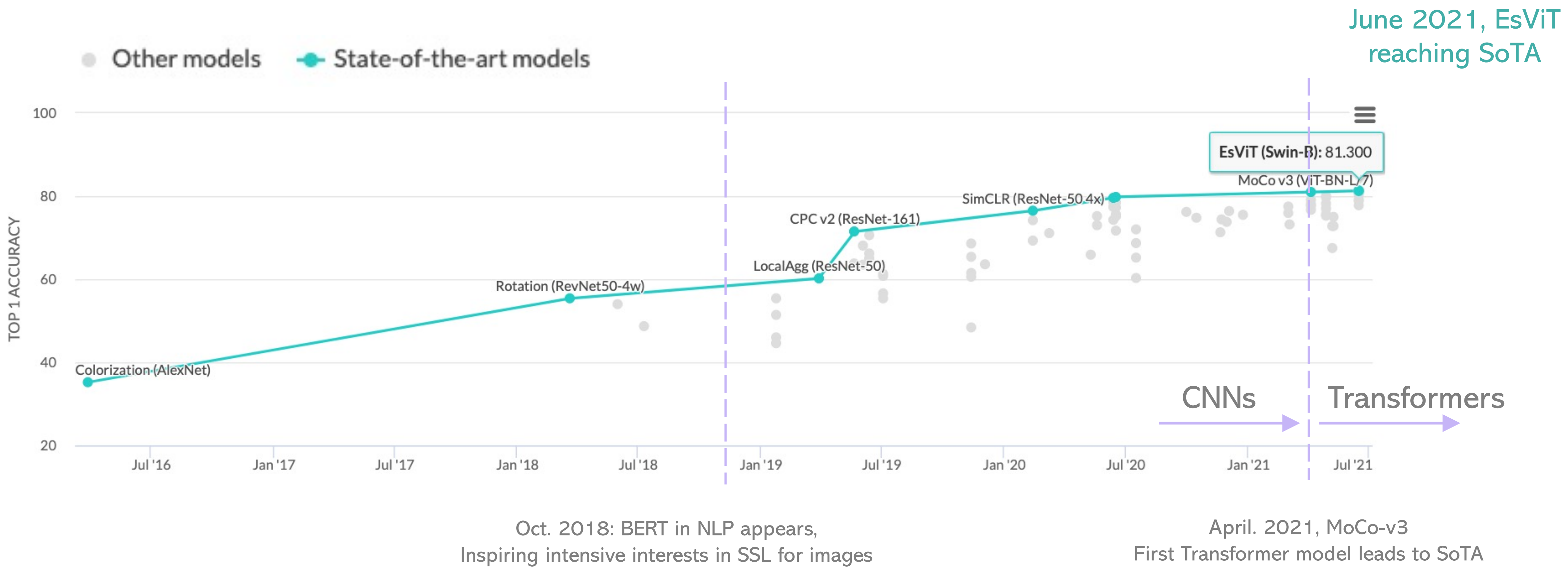
Fast

Automatic discovery of semantic correspondence between local regions





Self-Supervised Image Classification on ImageNet



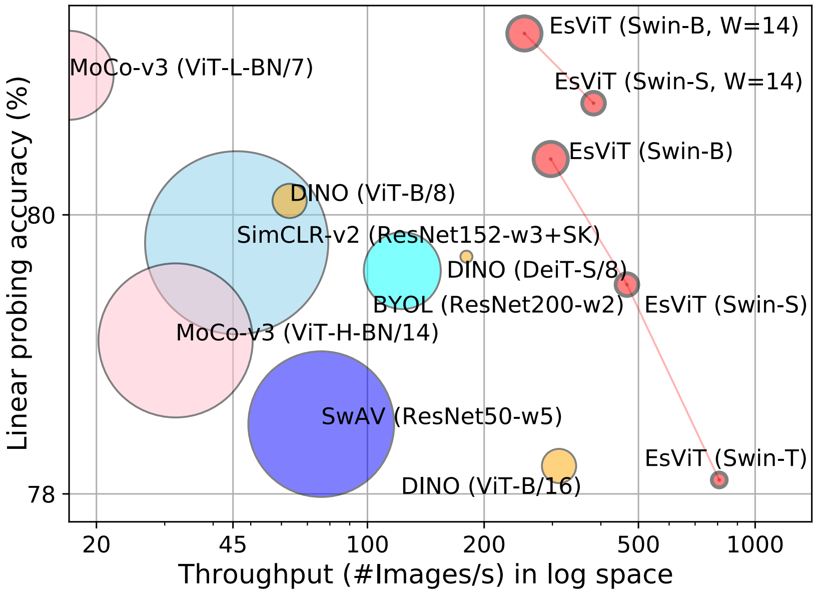
Source: https://paperswithcode.com/sota/self-supervised-image-classification-on

Efficiency *vs* Accuracy

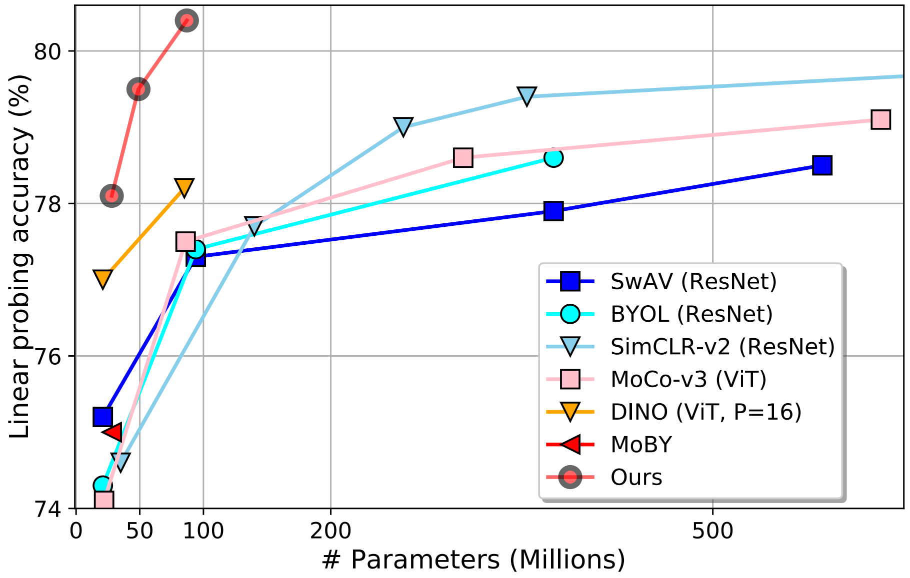
- 10x higher throughput, 3.5x smaller model size than prior arts
- Better scaling performance on accuracy vs. model size and throughput.

Model Size: 304 87 304 / 87 = 3.5x

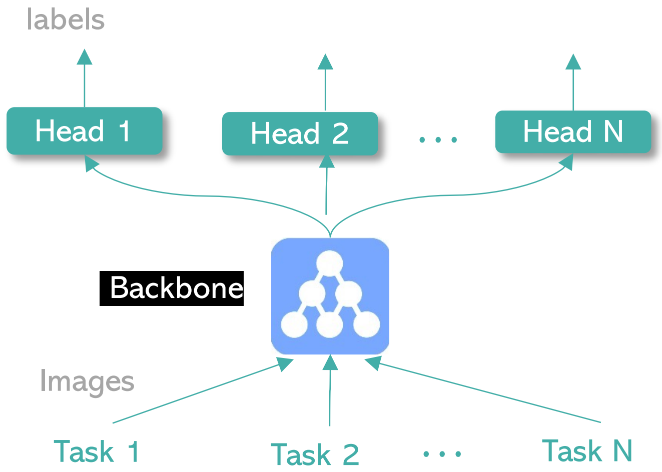
Throughput: 17 254 254 / 17 = 15x



Circle sizes indicates model parameter counts



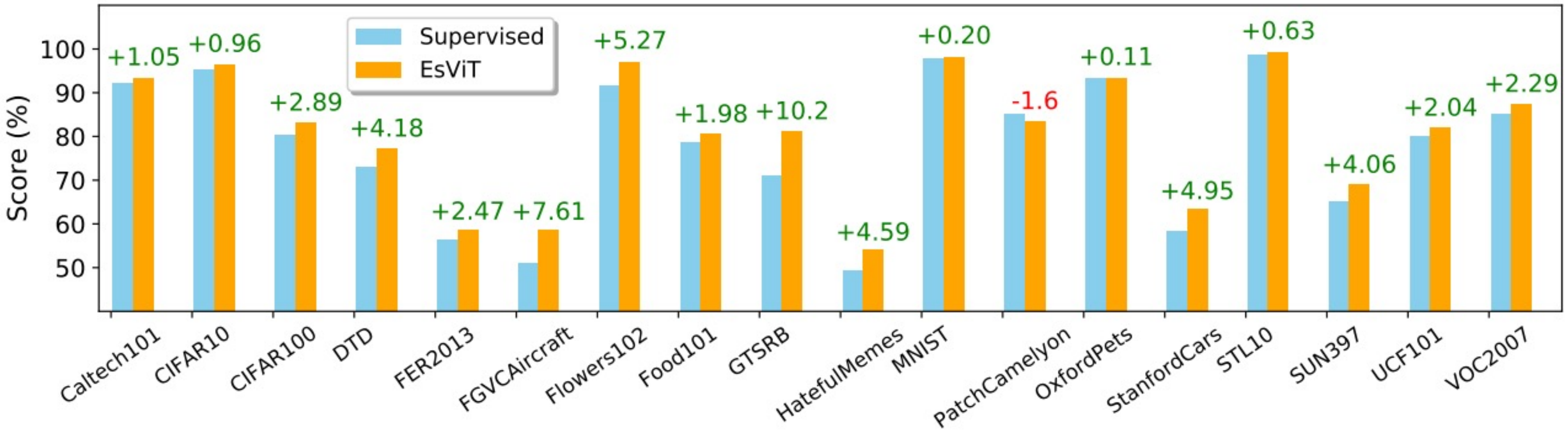
Transfer Learning



- Procedure: Pre-training a generic purpose **vision backbone**, and fine-tuning a **task-specific head** per task

  (Automatic hyper-parameter tuning is applied to ensure the comparison fairness)

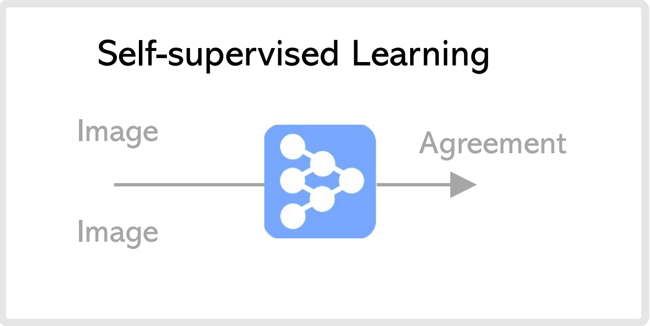
- EsViT outperforms the supervised counterpart on 17 out of 18 classification tasks

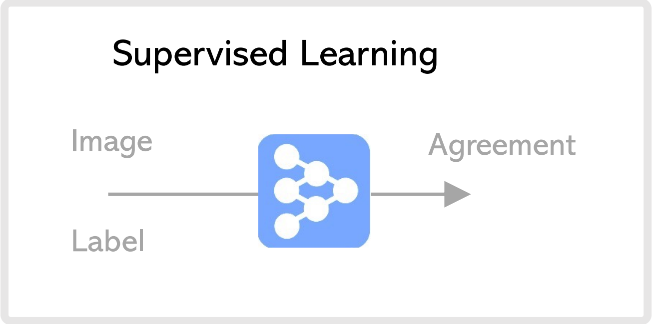


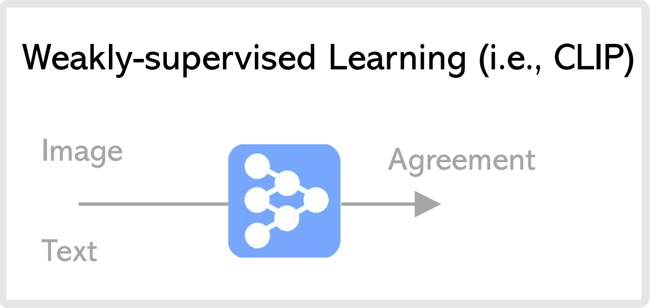
- Averaged scores: EsViT is comparable with CLIP, but uses 300x less pre-trained data

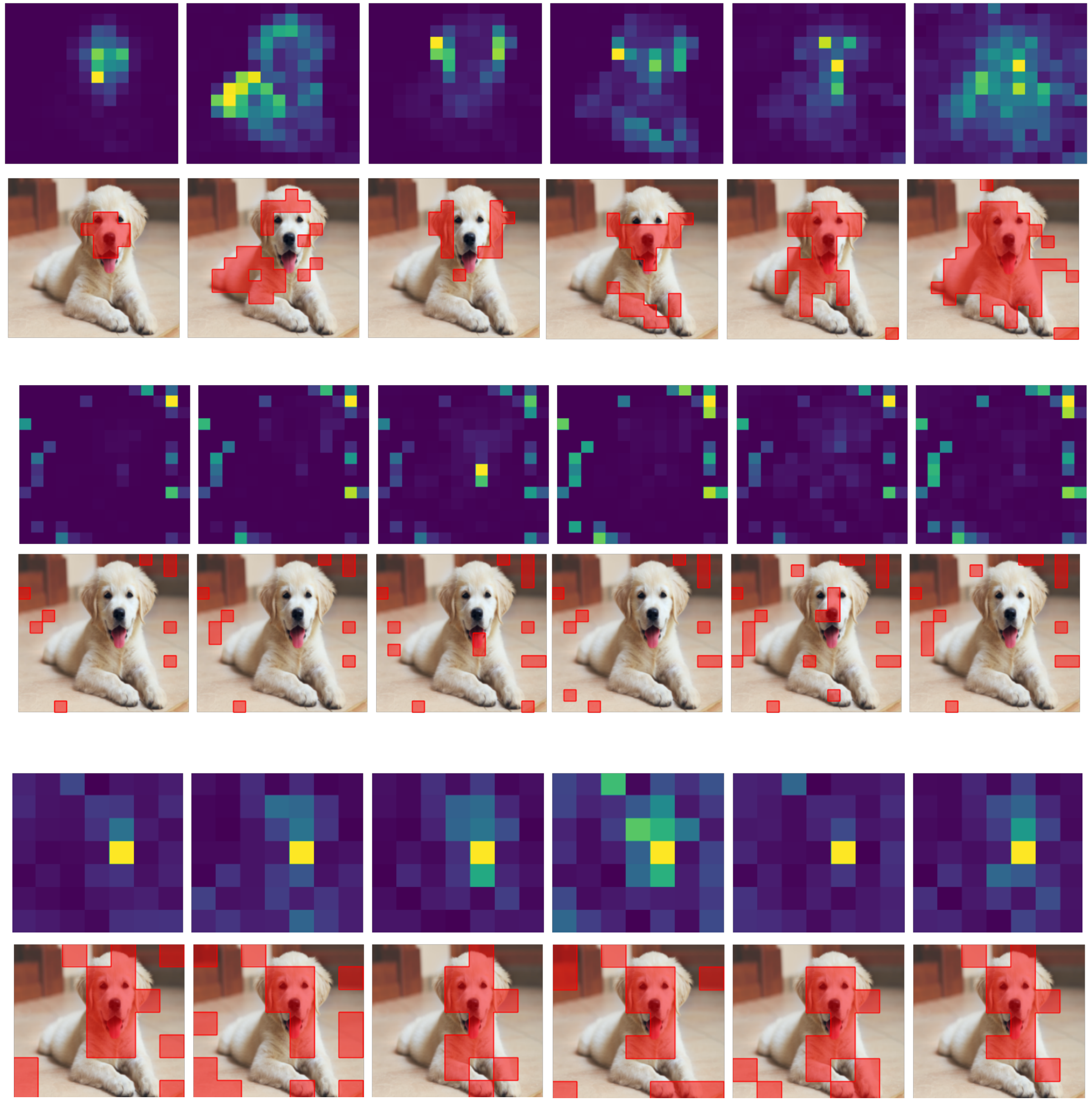
Method	Settings	Pre-training Data	Averaged Scores
EsViT	Self-supervised	1.2M images from ImageNet	**80.99**
Swin-T	Supervised	1.2M image-label pairs from ImageNet	77.29
CLIP	Weakly-supervised	400 M image-text pairs from web	80.86

Why does SSL generalize better than supervised learning?





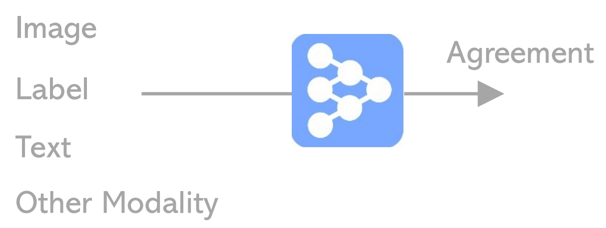




Summary & Future works

- **Future works**:
  - Generalizing EsViT to multi-modal learning, Each modality is considered as a view

A Unified Multimodal Learning Framework



- EsViT

Network architectures: A multi-stage Transformer architecture

Pre-training Objectives: A region-level pre-train task

GitHub https://github.com/microsoft/esvit

Personal Page: http://chunyuan.li

Thanks

Q & A

EsViT algorithm details

DINO updates teacher and student network alternatively: (*i*) Given a fixed teacher network, the student network is updated by minimizing the cross-entropy loss: $\boldsymbol{\theta}_s \leftarrow \arg\min_{\boldsymbol{\theta}_s} \mathcal{M}(s, t; \boldsymbol{\theta}_s)$, where $\mathcal{M}(s, t) = -p_t \log p_s$. (*ii*) The teacher model is updated as an exponential moving average (EMA) on the student weights $\boldsymbol{\theta}_t \leftarrow \lambda\boldsymbol{\theta}_t + (1 - \lambda)\boldsymbol{\theta}_s$, with λ following a cosine schedule from 0.996 to 1 during training. Please refer to [6] for details.

$$\mathcal{L}_V = \frac{1}{|\mathcal{P}|} \sum_{(s,t)\in\mathcal{P}} \mathcal{M}_V(s, t), \ \text{ with } \ \mathcal{M}_V(s, t) = -p_s \log p_t, \tag{1}$$

$$\mathcal{L}_R = \frac{1}{|\mathcal{P}|} \sum_{(s,t)\in\mathcal{P}} \mathcal{M}_R(s, t), \ \text{ with } \ \mathcal{M}_R(s, t) = -\frac{1}{T}\sum_{i=1}^{T} p_{j^*} \log p_i, \ \ j^* = \arg\max_j \frac{z_i^T z_j}{\|z_i\|\|z_j\|}, \tag{2}$$

Ablations on Networks Architectures and Pre-train Tasks

We briefly describe three schemes as follows, and benchmark them in the experiments. (*i*) *Swin Transformer* [39]: A shifted window partitioning approach is proposed, which alternates between two partitioning configurations in consecutive Transformer blocks, so that each local feature is grouped into different windows in self-attentions. (*ii*) *Vision Longformer (ViL)* [70]: Features in each local window are further allowed to attend all features in the 8-neighboring windows. (*iii*) *Convolution vision Transformer (CvT)* [62]: Features in neighboring windows are considered in the convolutional projection in self-attentions. Please refer each paper for detailed description.

Method	#Param.	Im./s	Pre-train tasks	Linear	k-NN
DeiT	21	1007	$\mathcal{L}_V$	75.9	73.2
ResNet-50	23	1237	$\mathcal{L}_V$	75.3†	67.5†
			$\mathcal{L}_V$	75.0	69.3
			$\mathcal{L}_V+\mathcal{L}_R$	**75.7**	**71.2**
Swin	28	808	$\mathcal{L}_V$	77.1	73.7
			$\mathcal{L}_V+\mathcal{L}_R$	**77.6**	**75.4**
ViL	28	386	$\mathcal{L}_V$	77.3	73.9
			$\mathcal{L}_V+\mathcal{L}_R$	**77.5**	**74.5**
CvT	29	848	$\mathcal{L}_V$	77.6	74.8
			$\mathcal{L}_V+\mathcal{L}_R$	**78.5**	**76.7**

Table 8: Different sparse attentions in EsViT with and without $\mathcal{L}_R$. DeiT and ResNet-50 are shown as references. † indicates numbers reported in [6].