Emerging Properties in Self-Supervised Vision Transformers

Дата публикации: 29 апреля 2021
Принята на ICCV 2021

Авторы:
Mathilde Caron [1,2] Hugo Touvron [1,3] Ishan Misra [1] Hervé Jégou [1]
Julien Mairal [2] Piotr Bojanowski [1] Armand Joulin [1]

[1] Facebook AI Research [2] Inria [3] Sorbonne University

У авторов есть различные публикации по темам image transformers, self supervision и contrasting clustering. DINO выглядит логичным продолжением предыдущих работ.

Цитирования: 97 по NASA ADS / 142 по Google Scholar / 159 по Semantic Scholar
36 Highly Influential Citations согласно Semantic Scholar

Статья во многом схожа с BYOL
Авторы признаются: *Our approach takes its inspiration from BYOL but operates with a different similarity matching loss and uses the exact same architecture for the student and the teacher.*

Есть подробное сравнение с конкурентами

Были побиты SOTA в self supervision, однако **не очевидна практическая ценность** на реальных задачах. Пора ли выбрасывать классические свёрточные сети?

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**