

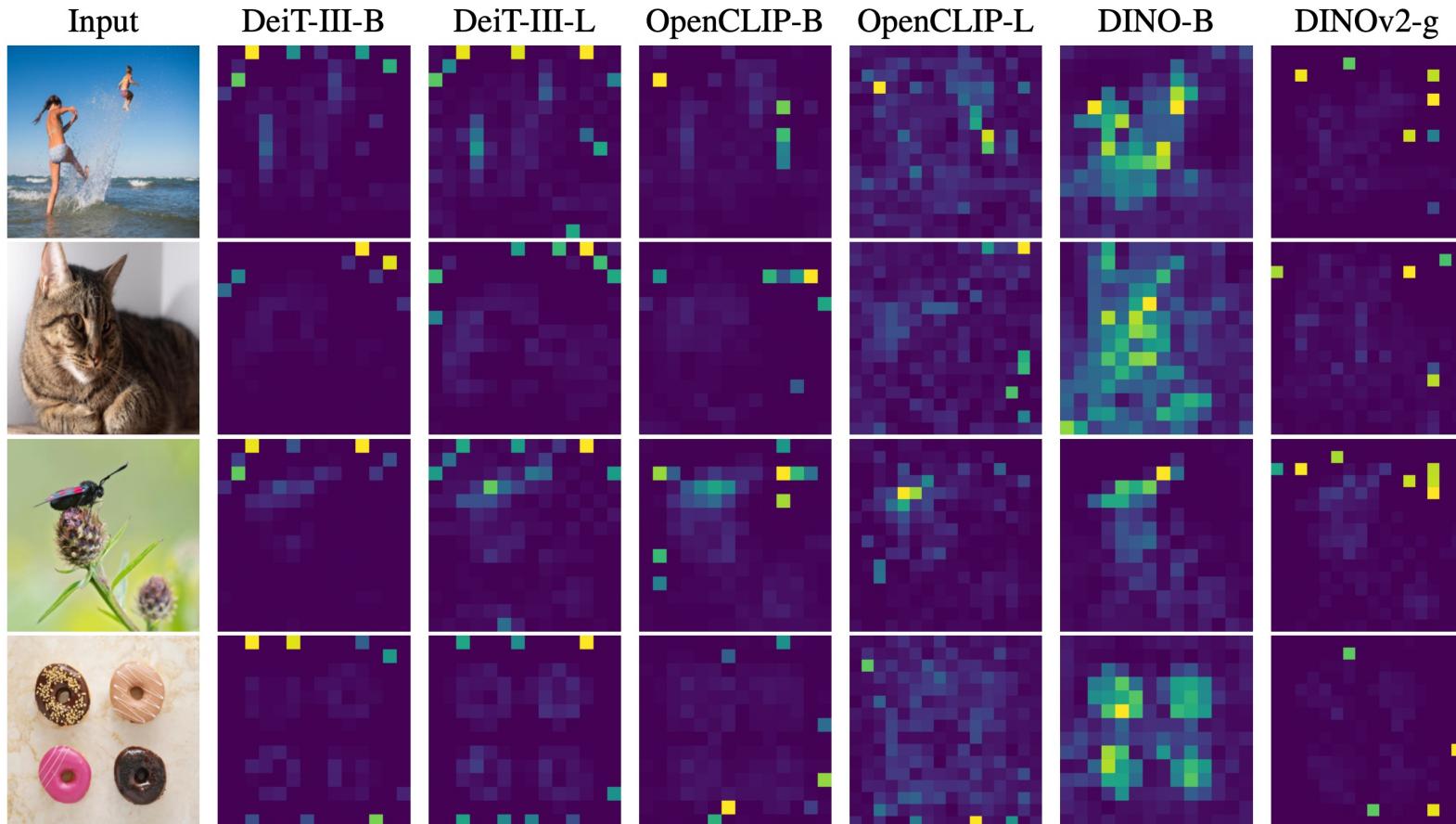
# Vision Transformers Need Registers



Яндекс

Egor Snigirev

# Что такое артефакты

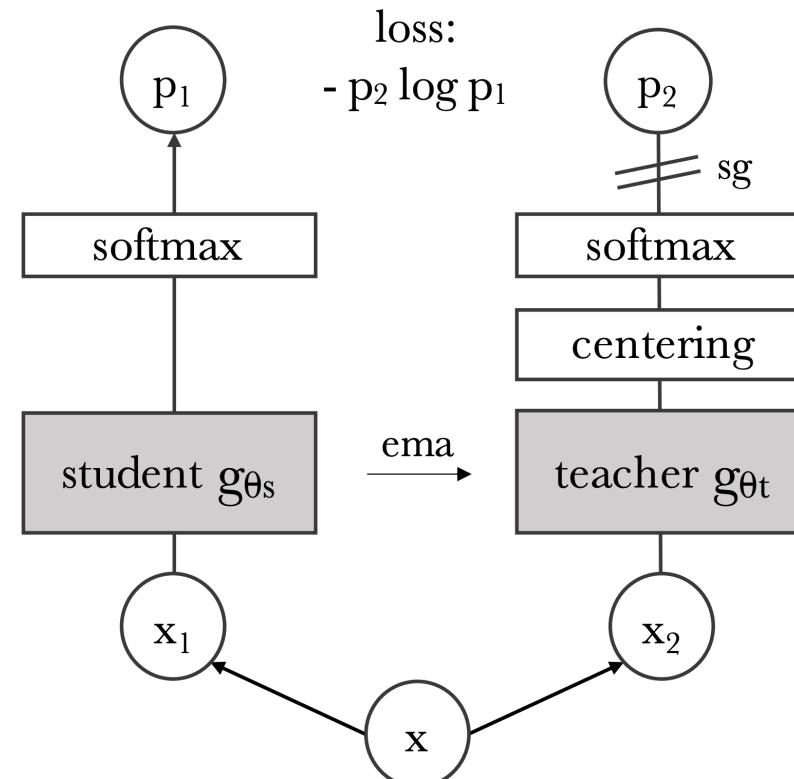
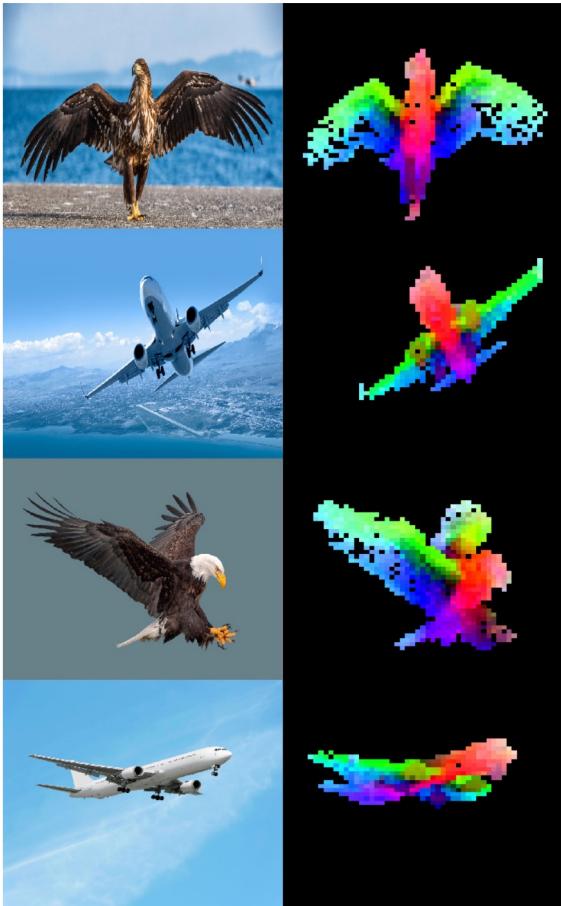


Attention maps современных трансформеров

# DINO



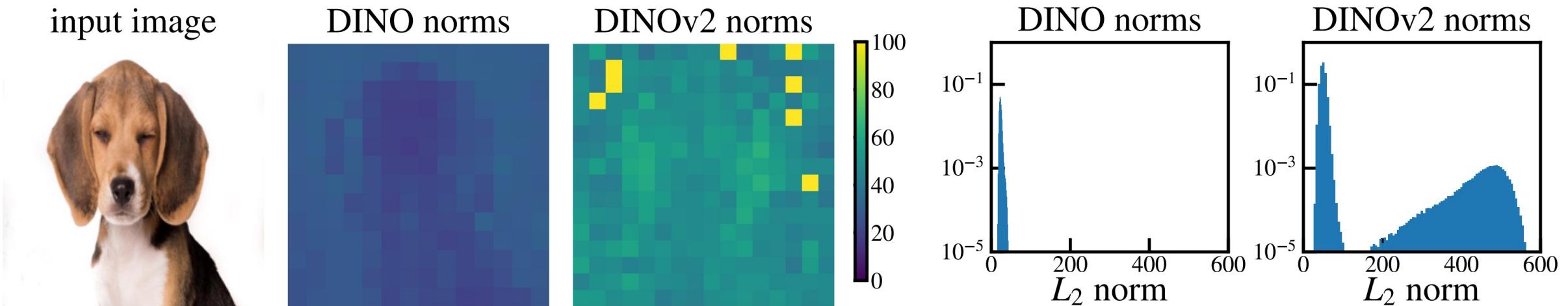
Self-supervised модель для выделения визуальных фичей



# Свойства артефактов



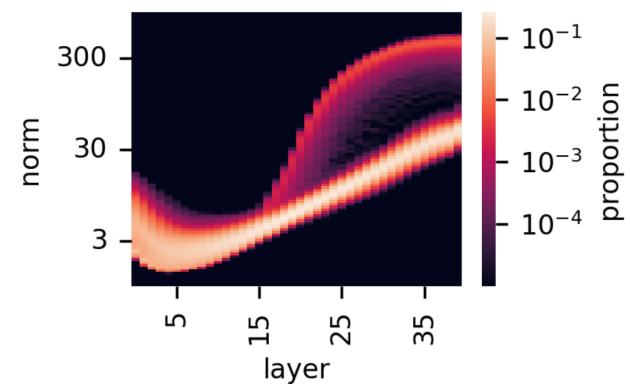
Артефакты – это токены с большой нормой



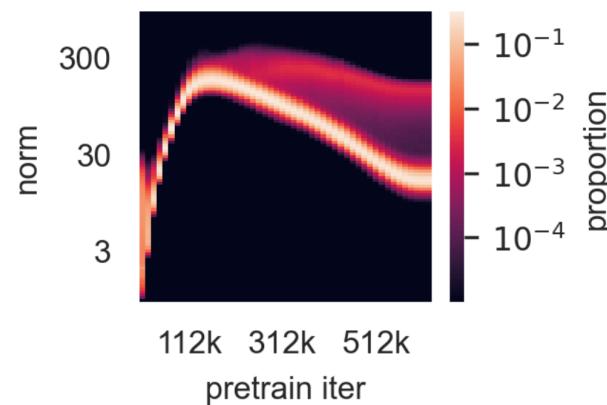
# Свойства артефактов



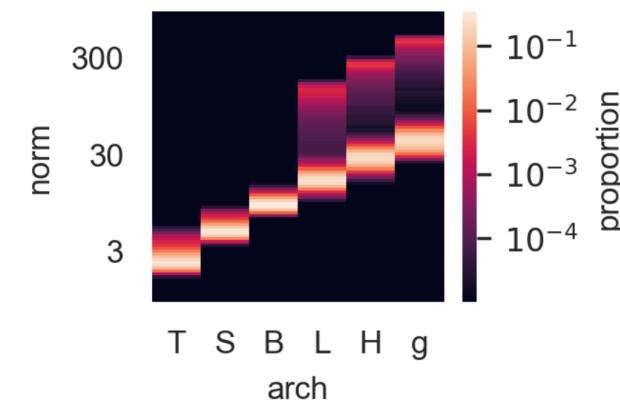
Артефакты появляются на больших моделях, на более поздних слоях и на более поздней стадии обучения



(a) Norms along layers.



(b) Norms along iterations.



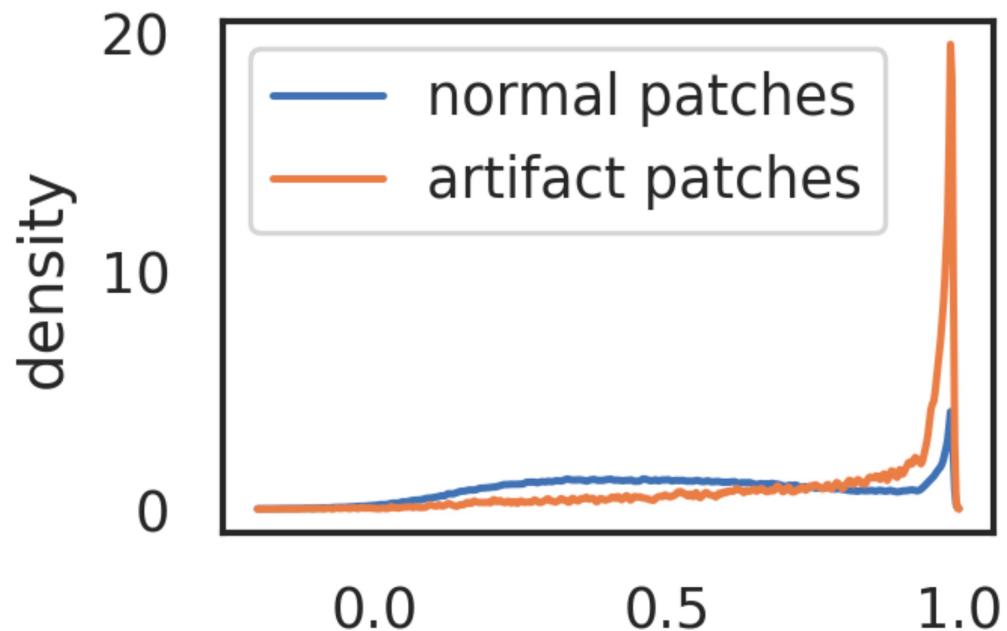
(c) Norms across model size.

Свойства артефактов для DINO ViT 40 слоев

# Свойства артефактов



Артефакты появляются на патчах с избыточной/ненужной информацией



Распределение косинусных расстояний от патчей до их соседей

# Свойства артефактов



В артефактах хранится мало локальной информации

	position prediction		reconstruction
	top-1 acc	avg. distance ↓	L2 error ↓
normal	<b>41.7</b>	<b>0.79</b>	<b>18.38</b>
outlier	22.8	5.09	25.23

Результаты линейных моделей, обученных предсказывать позицию патча и его пиксели

# Свойства артефактов



В артефактах хранится много глобальной информации

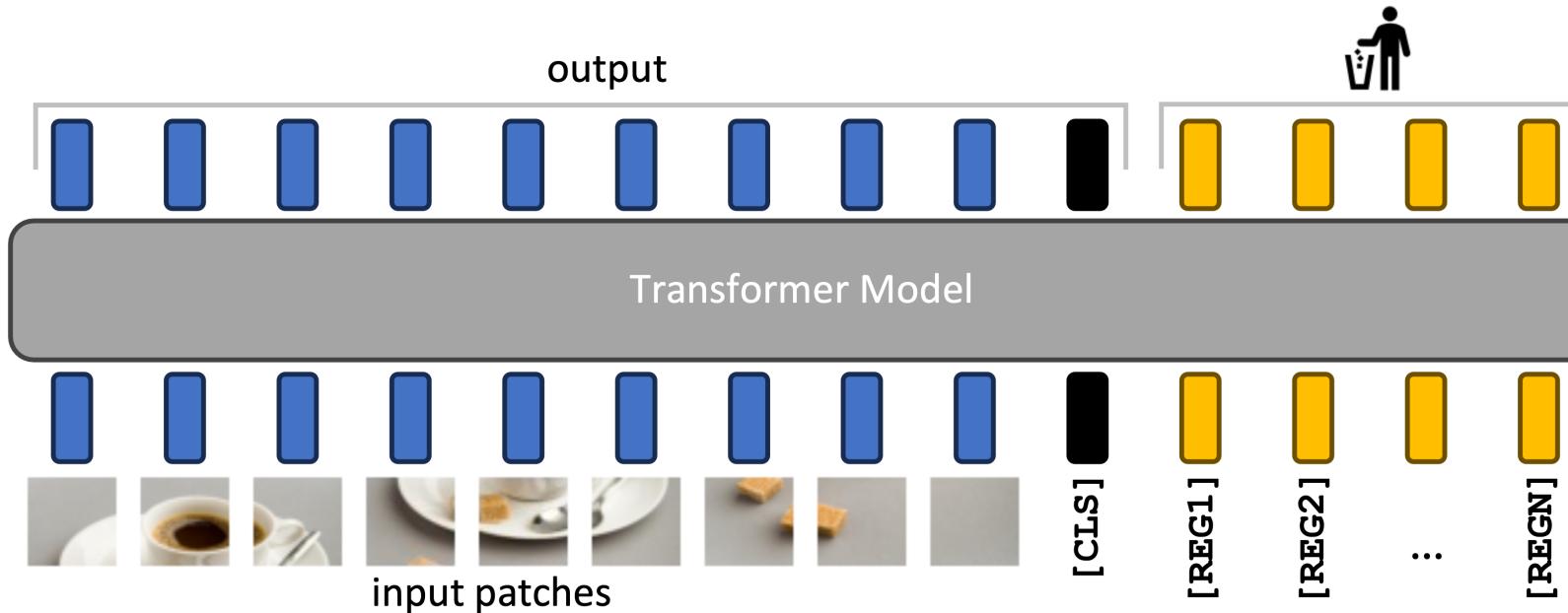
	IN1k	P205	Airc.	CF10	CF100	CUB	Cal101	Cars	DTD	Flow.	Food	Pets	SUN	VOC
[CLS]	<b>86.0</b>	<b>66.4</b>	<b>87.3</b>	<b>99.4</b>	<b>94.5</b>	<b>91.3</b>	<u>96.9</u>	<b>91.5</b>	<b>85.2</b>	<b>99.7</b>	<b>94.7</b>	<b>96.9</b>	<b>78.6</b>	<u>89.1</u>
normal	65.8	53.1	17.1	97.1	81.3	18.6	73.2	10.8	63.1	59.5	74.2	47.8	37.7	70.8
outlier	<u>69.0</u>	<u>55.1</u>	<u>79.1</u>	<u>99.3</u>	<u>93.7</u>	<u>84.9</u>	<b>97.6</b>	<u>85.2</u>	<u>84.9</u>	<u>99.6</u>	<u>93.5</u>	<u>94.1</u>	<u>78.5</u>	<b>89.7</b>

Результаты линейных моделей, обученных предсказывать позицию патча и его пиксели

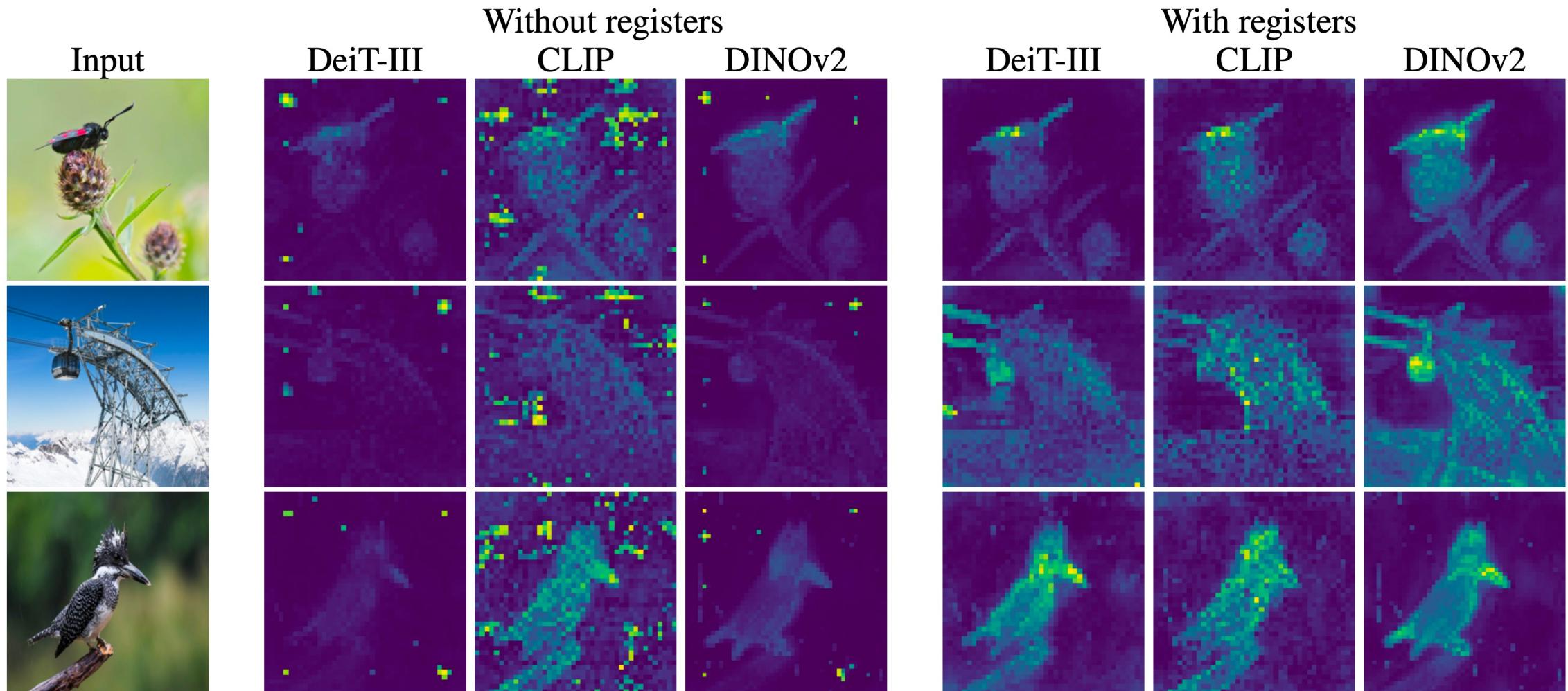
# Идея решения



Сделать отдельные токены, чтобы модели не пришлось складывать глобальную информацию в артефакты



# Экспериментальные результаты

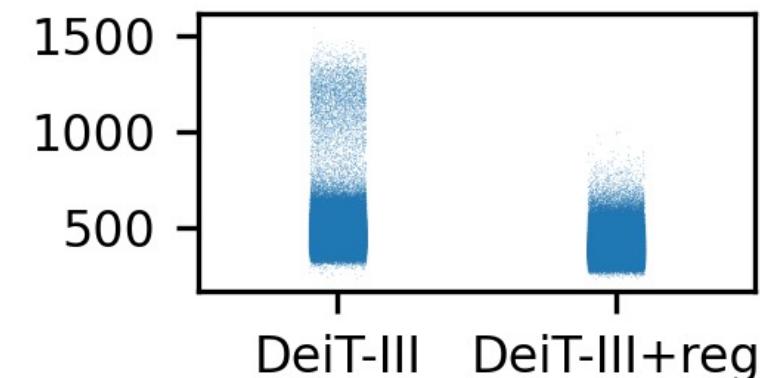
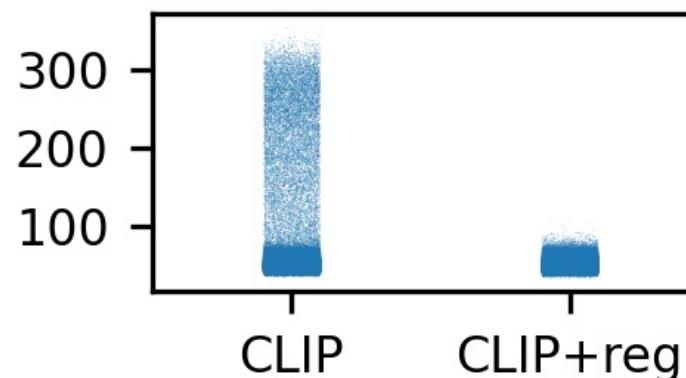
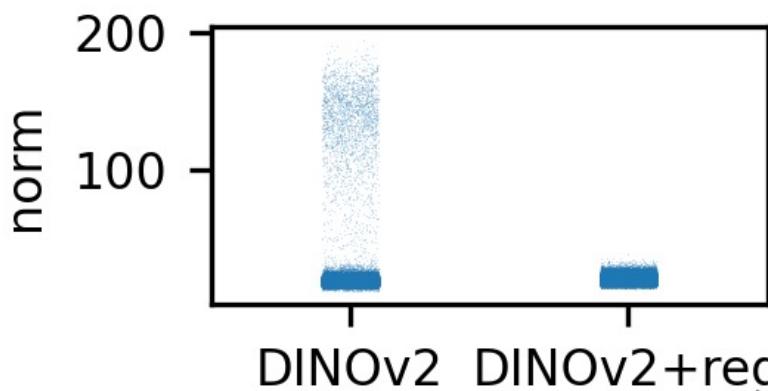


# Экспериментальные результаты



В экспериментах использовали методы

- DeiT-III - для классификации изображений
- OpenCLIP – для text-image моделей с энкодером ViT-B/16
- DINOv2 – self-supervised метод для построения фичей



С новыми регистрами пропали выбросы в нормах выхода

# Экспериментальные результаты

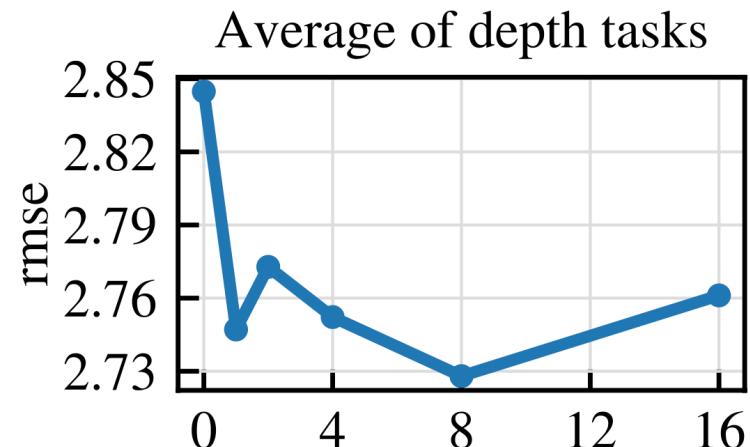
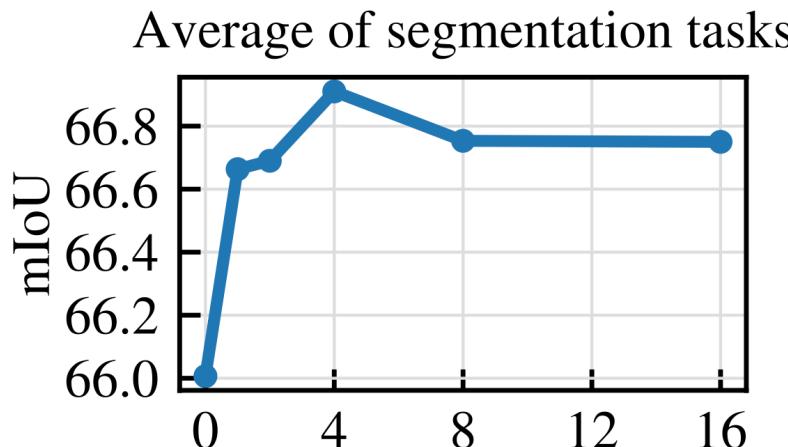
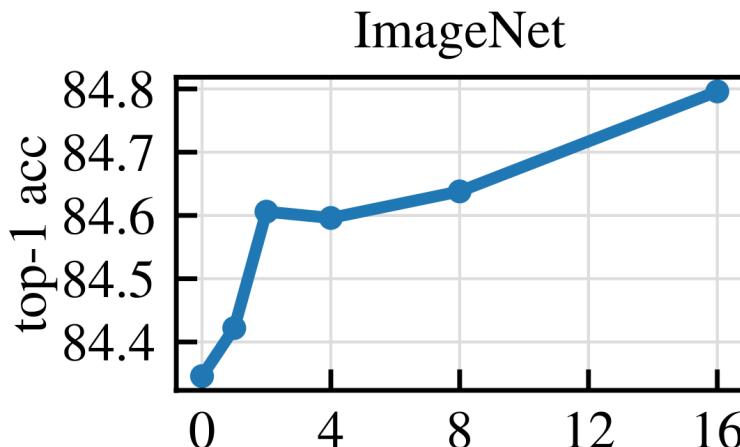


Влияние регистров на качество. Видно, что качество моделей не упало, а иногда даже немного улучшилось

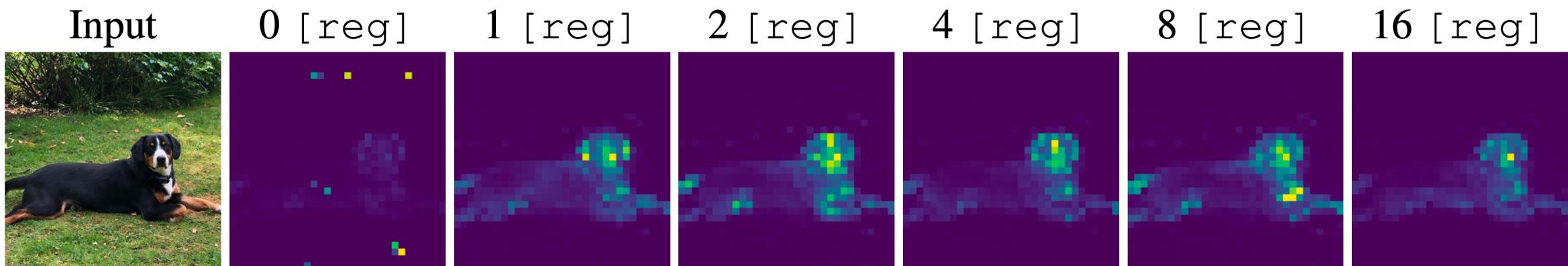
	Классификация ImageNet Top-1	Сегментация ADE20k mIoU	Глубина NYUD rmse ↓
DeiT-III	84.7	38.9	0.511
DeiT-III+reg	84.7	39.1	0.512
OpenCLIP	78.2	26.6	0.702
OpenCLIP+reg	78.1	26.7	0.661
DINOv2	84.3	46.6	0.378
DINOv2+reg	84.8	47.9	0.366

ImageNet Top-1
OpenCLIP
OpenCLIP+reg

# Экспериментальные результаты



Влияние количества регистров на качество



Влияние количества регистров на attention map

# Экспериментальные результаты

Решение задачи детекции без учителя. Обучается LOST на фичах, полученных из обученных моделей с регистрами и без

	VOC 2007	VOC 2012	COCO 20k
DeiT-III	11.7	13.1	10.7
DeiT-III+reg	27.1	32.7	25.1
OpenCLIP	38.8	44.3	31.0
OpenCLIP+reg	37.1	42.0	27.9
DINOv2	35.3	40.2	26.9
DINOv2+reg	55.4	60.0	42.0
DINOv1	61.9	64	50.7

# Attention maps для регистров

