Отчёт по проекту: SAE for CLIP

1 Методы обучения

Метрики и подходы взял из статьи [1]. На вход SAE подавал выходы из предпоследнего (11-го слоя) ViT-B/16. Обучал 3 версии SAE: на выходах блока self_attn, выходах mlp и выходах mlp с геометрической инициализацией смещения декодера (рис. 1). Параметры обучения: learning rate 4e-4, λ_{l1}=8e-5, expansion factor 64. Так же пробовал ReLU менять на TopK (k=32) (рис. 2). Модели обучал на трейн части репака ImageNet-1k с обрезкой по центру до квадратной формы и масштабированием до 256x256 [2]. Статистики считал на тренировочных и валидационных данных (рис. 3).

Для визуализации результатов усреднял активации нейронов по изображениям и по патчам. Приложение загружает готовые агрегации в виде топ 16 изображений (активаций) на каждый нейрон и маски из патчей, входящих в топ активаций на выбранный нейрон.

2 Результаты

Получить распределение как в статье не получилось: в целом кластеры есть, но выглядят совсем иначе (рис. 3).

На выходах self_attn вектора получились более разряженными (рис. 4), но проигрывают в информативности: число нейронов, которые активируются на различные изображения, в разы меньше по сравнению с обучением на выходах mlp (рис. 5).

При обучении с ReLU небольшое число нейронов активируются на все 50000 изображений из валидационной выборки (на выходах self_attn 52, на выходах mlp 429, на выходах mlp с геометрической инициализацией 418), остальные — на небольшое число картинок (по бóльшей части на 1-2 картинки). При обучении с TopK такой зависимости нет (рис. 5).

Для отрисовки сегментаций усреднял активации по патчам и классам (рис. 6). Кажется, что для хорошей маски нужна более умная группировка по какому-то признаку.

Таблица 1: Параметры и метрики SAE на валидационных данных

№	CLIP output	Activation	Geom. dec. bias	MSE	Mean activation	L0	Entropy
1	self_attn	ReLU	False	0.0068	0.2844	0.0010	8649.3
2	self_attn	TopK	False	0.0030	0.4721	0.0002	8606.5
3	mlp	ReLU	False	0.0188	0.2909	0.0086	8660.5
4	mlp	TopK	False	0.0149	1.0397	0.0002	8646.7
5	mlp	ReLU	True	0.0188	0.2976	0.0084	8660.5

Список литературы

[1] Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

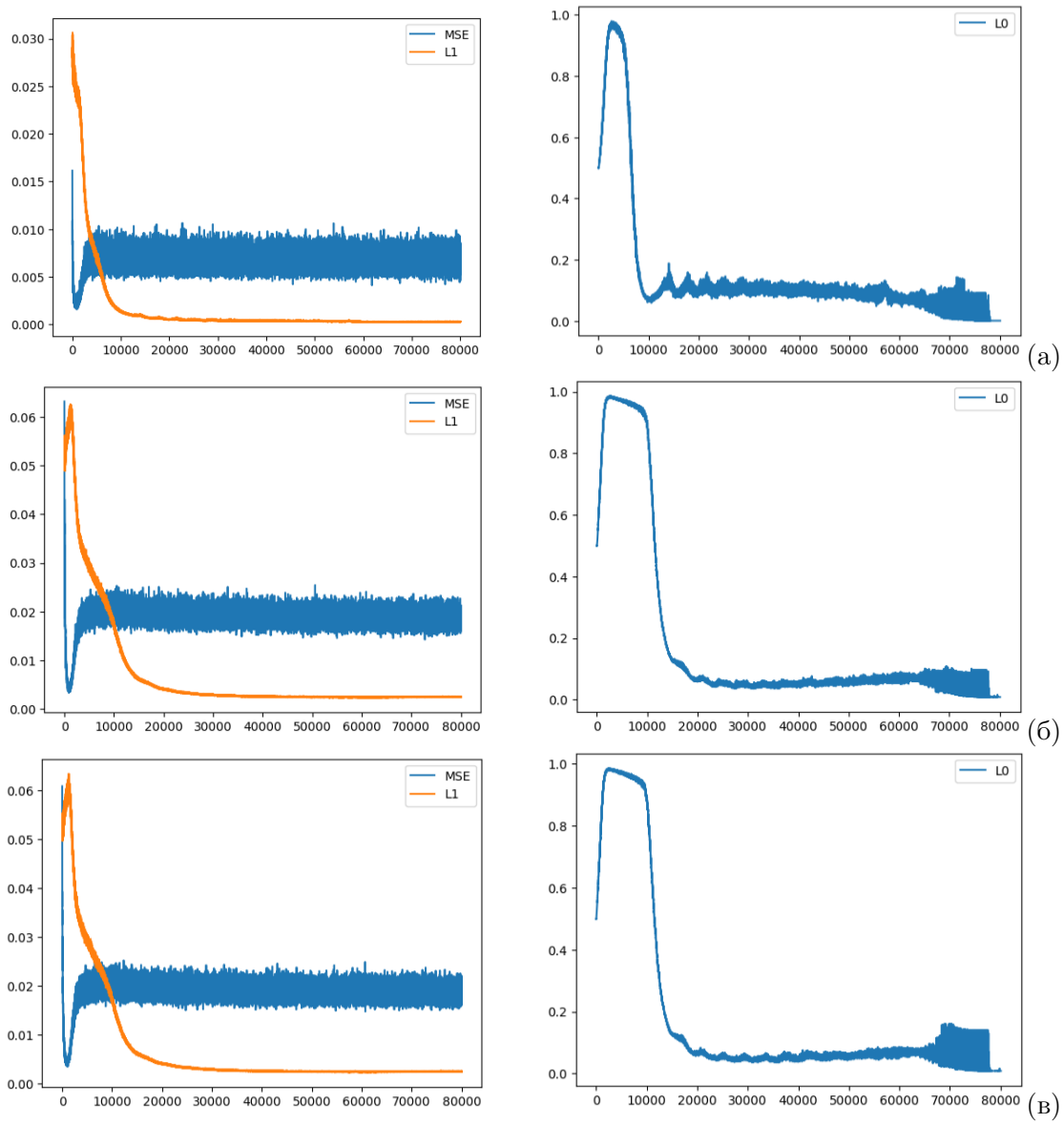

Рис. 1: Кривые обучения SAE с активацией ReLU на выходах self_attn (а), mlp (б), mlp с геометрической инициализацией (в).

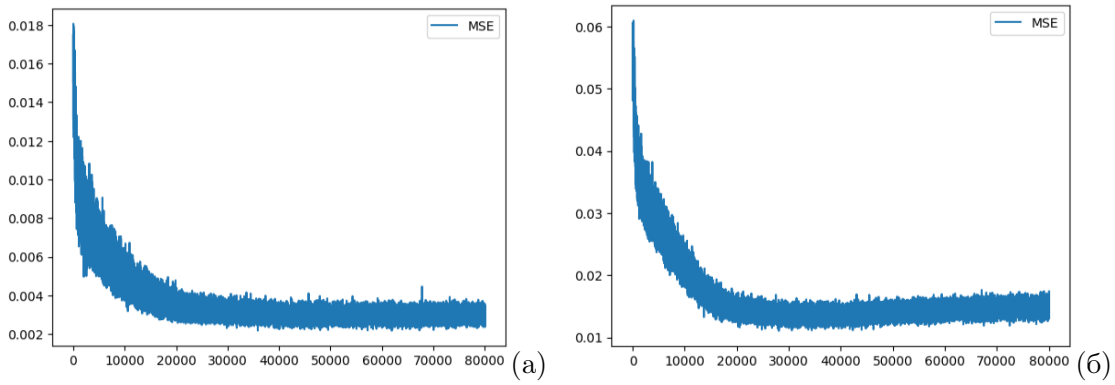

Рис. 2: Кривые обучения SAE с активацией TopK на выходах self_attn (а) и mlp (б).

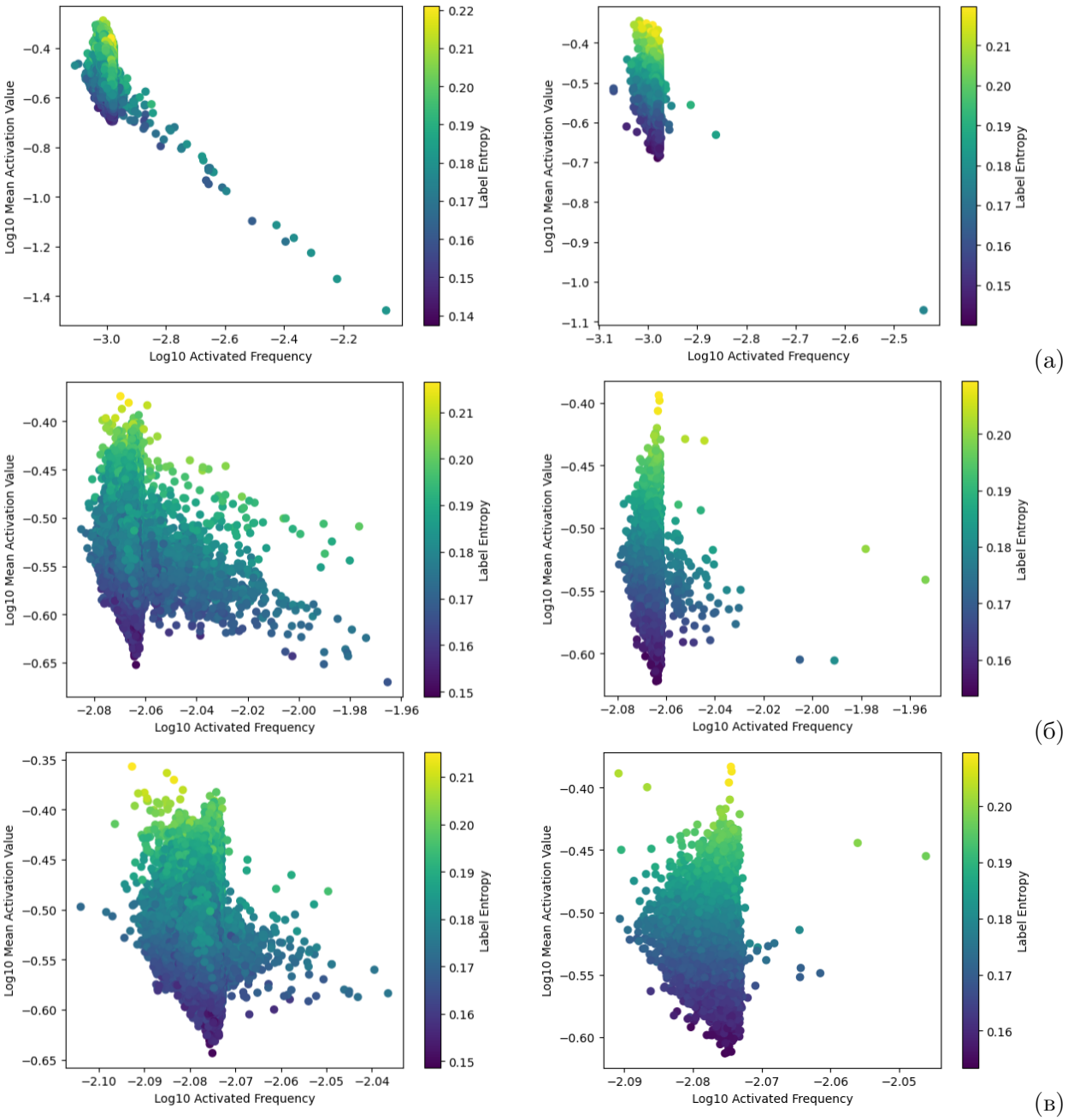

Рис. 3: Метрики латентных векторов SAE с активацией ReLU на выходах self_attn (а), mlp (б), mlp с геометрической инициализацией (в). Слева тренировочные данные, справа валидационные данные.

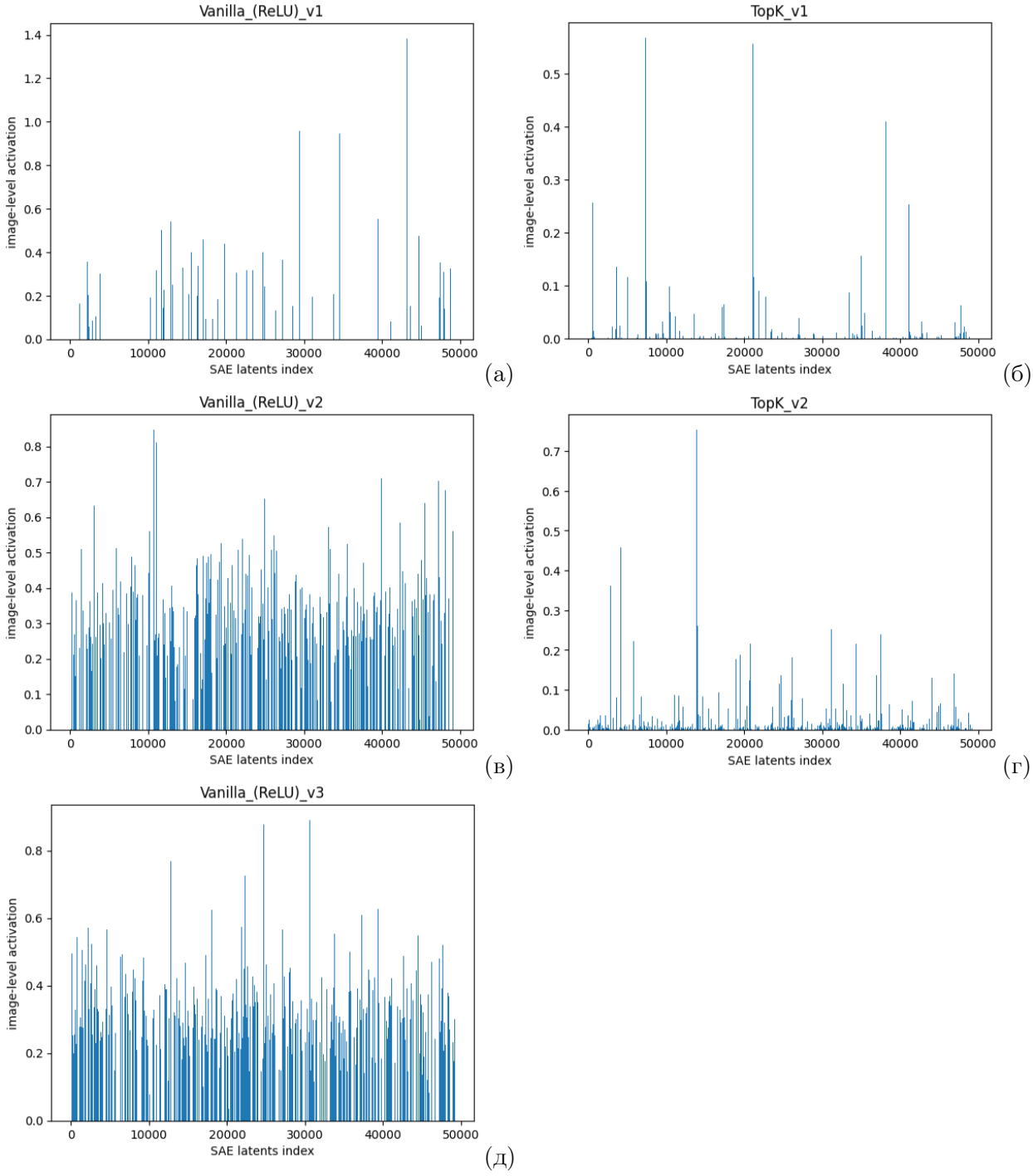

Рис. 4: Активации латентных векторов SAE на выходах self_attn (а, б), mlp (в, г), mlp с геометрической инициализацией (д). Функция активации ReLU слева, TopK справа.

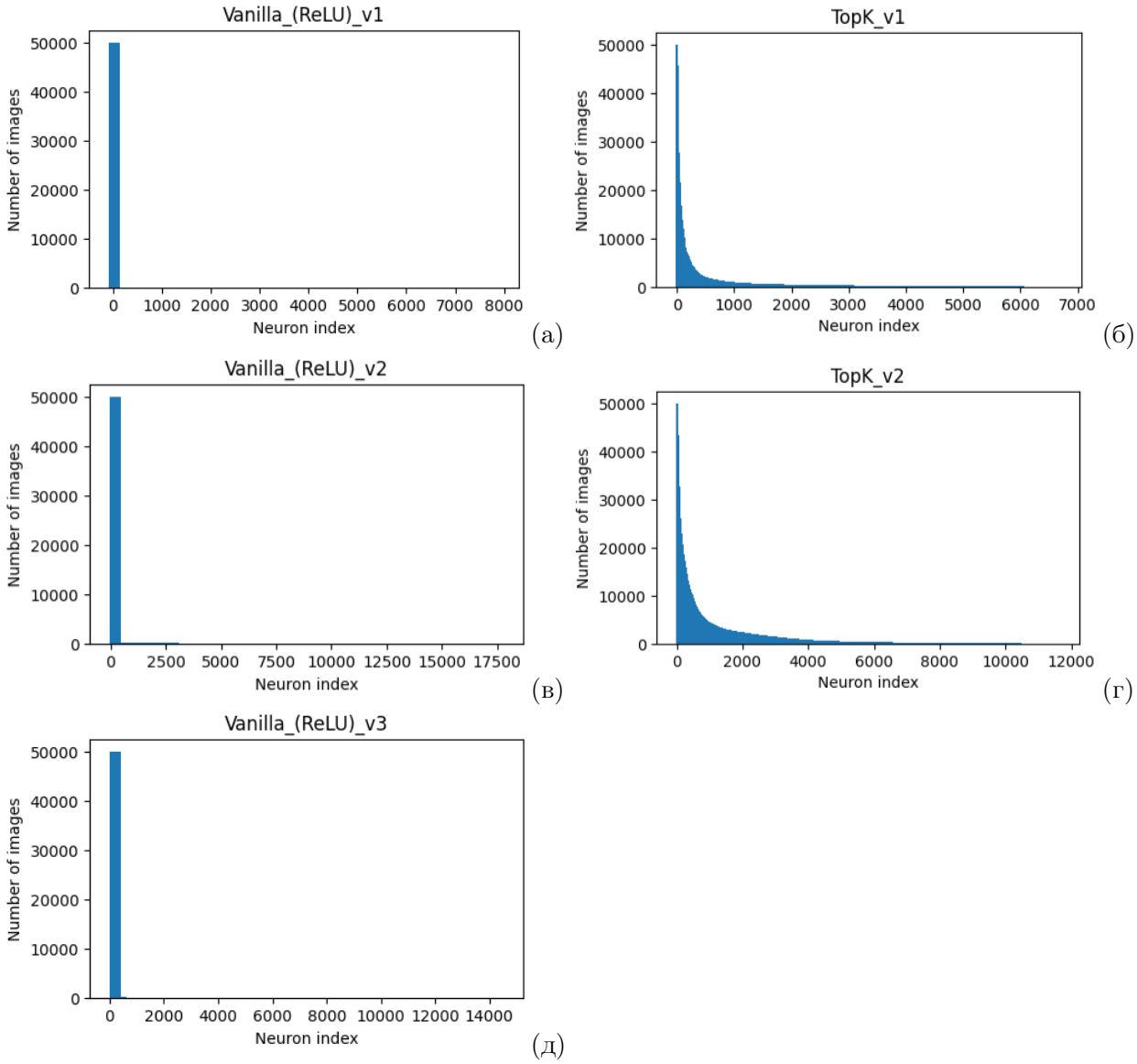

Рис. 5: Число активаций (изображений) на нейроны латентных векторов SAE на выходах self_attn (а, б), mlp (в, г), mlp с геометрической инициализацией (д). Функция активации ReLU слева, TopK справа.

Рис. 6: Маска активации нейрона 22296 на класс «tabby, tabby cat» модели SAE на выходах mlp с геометрической инициализацией.