Контрастное обучение в задачах компьютерного зрения для повышения интерпретируемости модели

Андрей Семёнов[124] Владимир Иванов[3] Александр Безносиков[124]

[1]MIPT [2]Yandex Research [3]Innopolis University [4]MMO Laboratory

Май 2024

Sparse Concept Bottleneck Models: Gumbel Tricks in Contrastive Learning
https://arxiv.org/abs/2404.03323
ICML 2024 submit

Содержание

Concept Bottleneck модели

$y \in \mathbb{R}$, $x \in \mathbb{R}^{\mathrm{d}}$, $c \in \mathbb{R}^{\mathrm{k}}$, $\{(x^{(i)}, y^{(i)}, c^{(i)})\}_{i=1}^{n}$,
$g : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{k}}$, $F : \mathbb{R}^{\mathrm{k}} \to \mathbb{R}$, (F, g) – информационный боттлнек

- Independent bottleneck $\hat{F} = \underset{F}{\operatorname{argmin}} \sum_i \mathcal{L}_{\mathrm{Y}}(F(c^{(i)}), y^{(i)})$

  $\hat{g} = \underset{g}{\operatorname{argmin}} \sum_{i,j} \mathcal{L}_{\mathrm{C_j}}(g_{\mathrm{j}}(x^{(i)}))$
- Sequential bottleneck $\hat{F} = \underset{F}{\operatorname{argmin}} \sum_i \mathcal{L}_{\mathrm{Y}}(F(\hat{g}(x^{(i)})), y^{(i)})$
- Joint bottleneck
  $\hat{F}, \hat{g} = \underset{F,g}{\operatorname{argmin}} \sum_i \left[\lambda_1 \mathcal{L}_{\mathrm{Y}}(F(g(x^{(i)}))) + \lambda_2 \sum_j \mathcal{L}_{\mathrm{C_j}}(g_{\mathrm{j}}(x^{(i)}))\right]$
- Standard model $\hat{F}, \hat{g} = \underset{F,g}{\operatorname{argmin}} \sum_i \mathcal{L}_{\mathrm{Y}}(F(g(x^{(i)})), y^{(i)})$

Сопутствующие работы

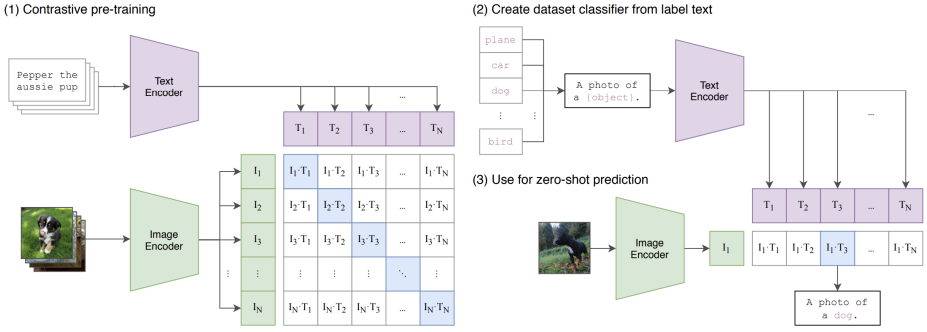


Рис.: CLIP

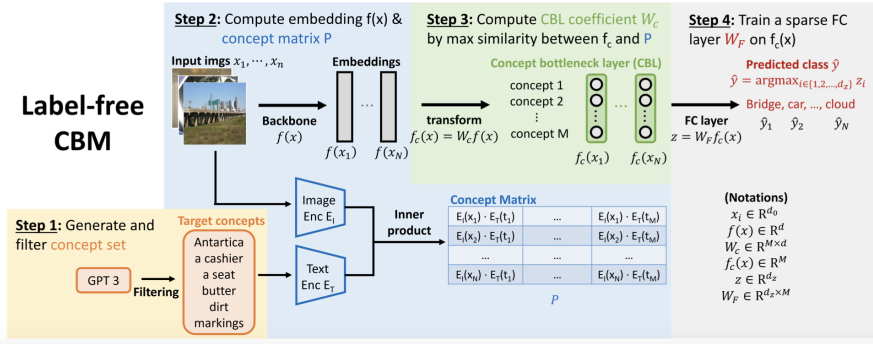


Рис.: Label-free

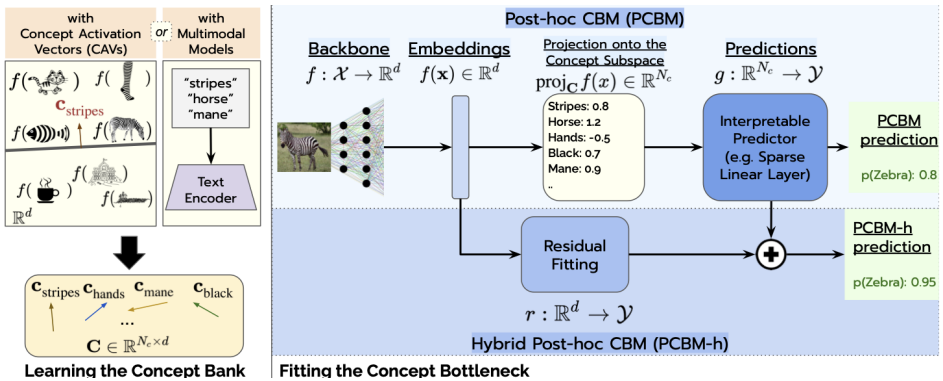


Рис.: Post-hoc CBM

Наш вклад

- Мы предложили два новых варианта архитектуры и алгорим для обучения Concept Bottleneck моделей (CBM).
- Мы формально описываем алгоритм для повышения точности CLIP – Concept Matrix Search (CMS), и в то же время делаем модель более интерпретируемой. Мы также приводим анализ латентного пространства CLIP.
- Получили неожиданный результат о полезности внутренних разреженных слоев в CBM.
- Исследовали задачу распределенной оптимизации, предложили способы устранения распределений тяжелых шумов в совместной задаче NLP и CV.

Градиентный клиппинг

SGD:

$$x^{k+1} = x^k - \gamma \nabla f(x^k, \xi^k) \tag{1}$$

Clipped-SGD:

$$x^{k+1} = x^k - \gamma \texttt{clip}(\nabla f(x^k, \xi^k), \lambda) \tag{2}$$

- $\texttt{clip}(x, \lambda) = \min\{1, \lambda / \|x\|\} x$
- $\mathbb{E}_{\xi^k}[\texttt{clip}(\nabla f(x^k, \lambda))] \neq \nabla f(x^k)$
- При $\beta_1 = 0$ Adam можно интерпретировать как Clipped-SGD с "адаптивным" λ

Тяжелый шум: теоретическая справка

Случайный вектор X имеет распределение с легкими хвостами если:

$$\mathbb{P}\{\|X - \mathbb{E}[X]\|\} \leq 2\exp\left(-\frac{b^2}{2\sigma^2}\right) \forall b > 0, \tag{3}$$

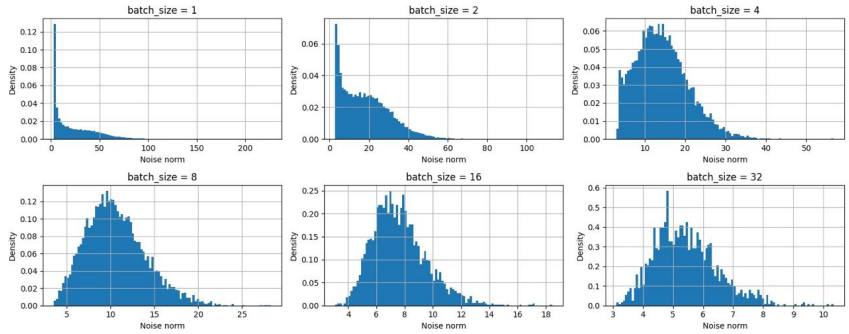
что эквивалентно

$$\mathbb{E}\left[\exp\left(\frac{\|X - \mathbb{E}[X]\|^2}{\sigma^2}\right)\right] \leq \exp 1 \tag{4}$$

Для задач с шумом в виде тяжелых хвостов в теории используют предположение:

$$\forall \alpha \in (1, 2] : \mathbb{E}[\|X - \mathbb{E}[X]\|^\alpha] \leq \sigma^\alpha \tag{5}$$

Тяжелый шум: экспериментальное подтверждение



Нормы градиентов, CIFAR10.

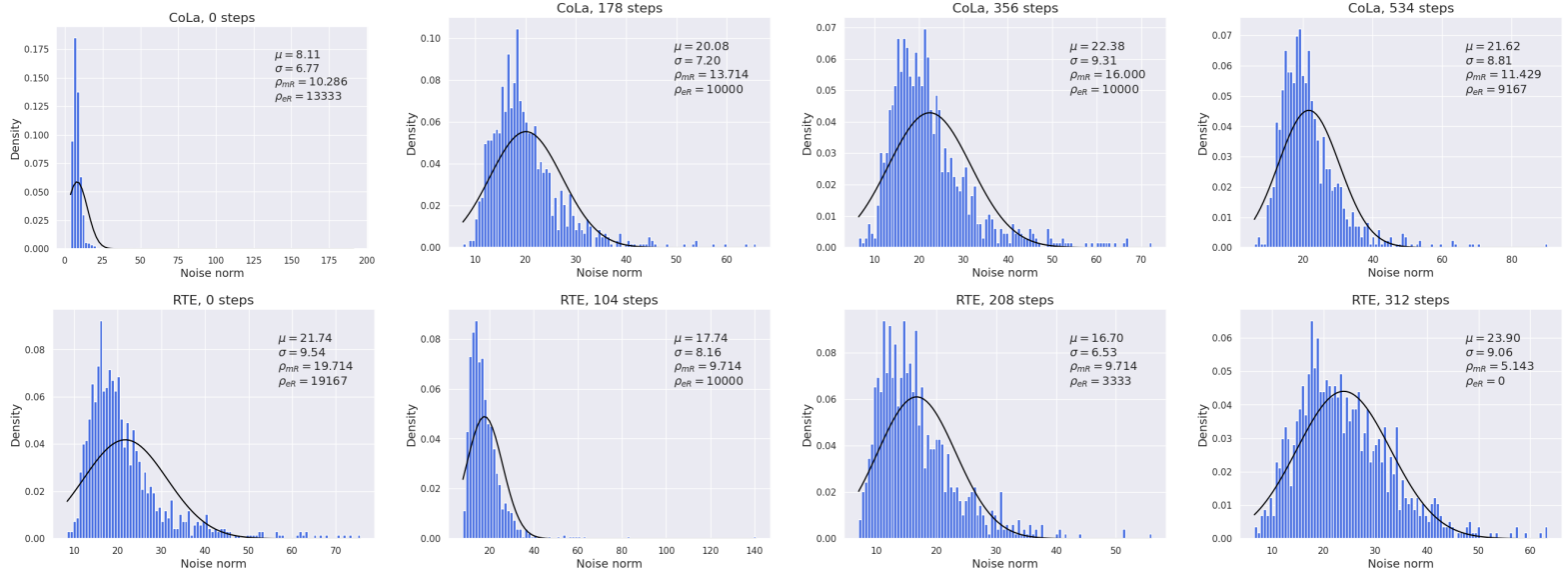
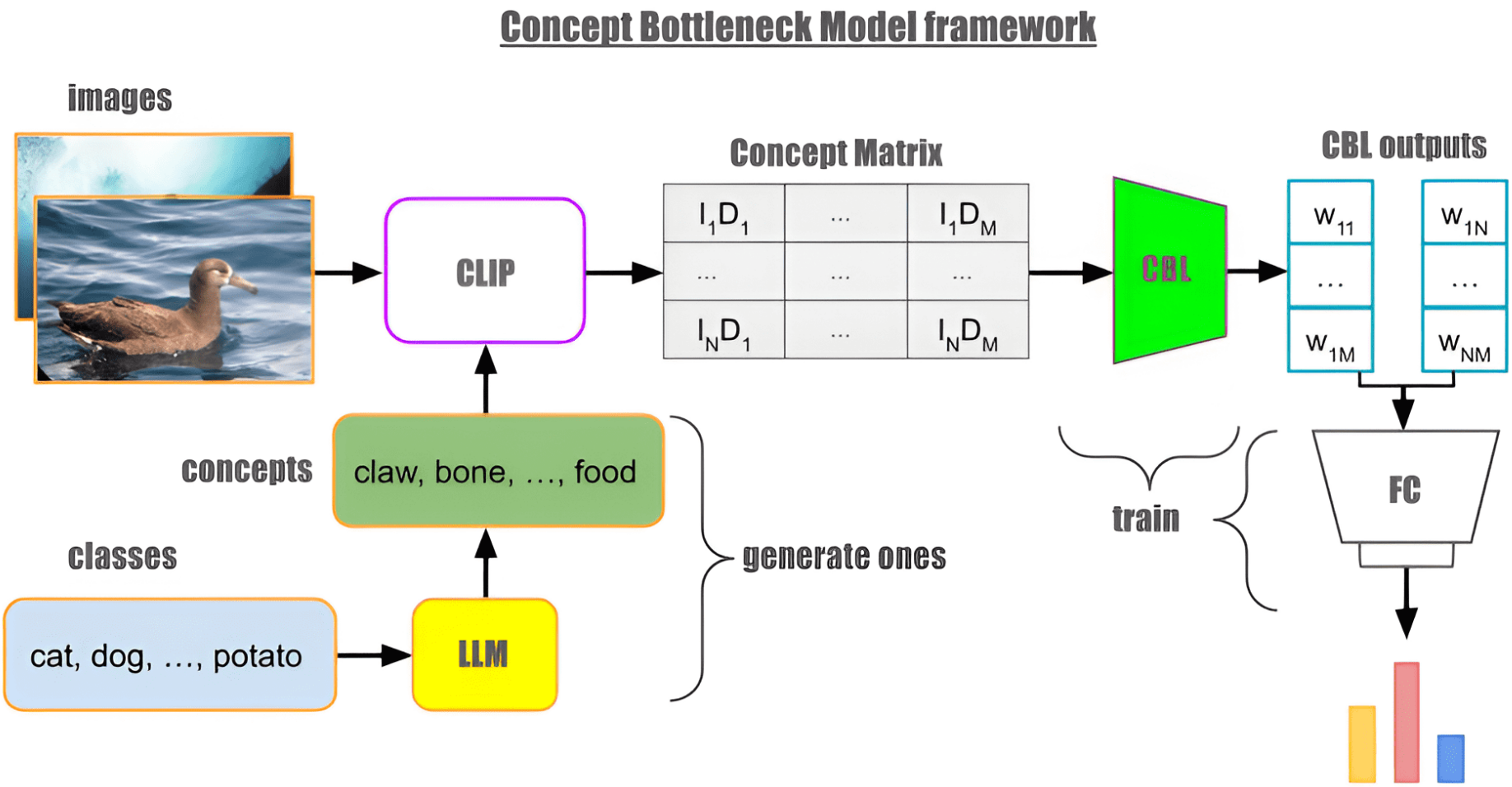


Рис.: Оценка градиентного шума для `Adam` на датасетах CoLa (первая строка) и RTE (вторая строка).

Наша архитетура для CBM



Наш CBM фреймворк.

Результаты: Общая схема для наших CBM

- i-эмбеддинг картинки, d-эмбеддинг концепта, $\mathcal{D} = (x, t, l)$ – датасет.
- CLIP: $\psi(x, t) = \left(\langle i, d_1 \rangle, \ldots, \langle i, d_{|\mathrm{D}|} \rangle\right)^\top \in \mathbb{R}^{|\mathrm{D}|}$

$$\min_{W_{\mathrm{CBL}}} \mathop{\mathbb{E}}_{(x,t,l)\sim\mathcal{D}} \left[\mathcal{L}_{\mathrm{CBL}}(W_{\mathrm{CBL}}\psi(x, t))\right], \tag{6}$$

$$\min_{W_{\mathrm{F}}} \mathop{\mathbb{E}}_{(x,t,l)\sim\mathcal{D}} \left[\mathcal{L}_{\mathrm{CE}}(W_{\mathrm{F}} W_{\mathrm{CBL}}\psi(x, t), l)\right]. \tag{7}$$

Результаты: Contrastive-CBM

Contrastive-CBM:

$$-\frac{1}{2|\mathrm{B}|}\sum_{k=1}^{|\mathrm{B}|}\left(\log\frac{e^{\alpha\langle w_k,\varphi_k\rangle}}{\sum_{j=1}^{|\mathrm{B}|}e^{\alpha\langle w_k,\varphi_j\rangle}}+\log\frac{e^{\alpha\langle w_k,\varphi_k\rangle}}{\sum_{j=1}^{|\mathrm{B}|}e^{\alpha\langle w_j,\varphi_k\rangle}}\right). \tag{8}$$

Результаты: Sparse-CBM

Sparse-CBM:

$$z = 1\left(\arg\max_{k}\left[g_k + \log \pi_k\right]\right), \tag{9}$$

$$-\frac{1}{2|B|}\sum_{k=1}^{|B|}\left(\log\frac{e^{(\log(\alpha\langle w_k,\varphi_k\rangle)+g_k)/\tau}}{\sum_{j=1}^{|B|}e^{\left(\log(\alpha\langle w_k,\varphi_j\rangle)+g_j\right)/\tau}} + \log\frac{e^{(\log(\alpha\langle w_k,\varphi_k\rangle)+g_k)/\tau}}{\sum_{j=1}^{|B|}e^{\left(\log(\alpha\langle w_j,\varphi_k\rangle)+g_j\right)/\tau}}\right). \tag{10}$$

$u \in \mathrm{Uniform}(0,1),\ g = -\log\log u \in \mathrm{Gumbel}(0,1)$

Результаты: ℓ_1-CBM

ℓ_1-CBM:

$$\min_{W_{\text{CBL}}} \mathop{\mathbb{E}}_{(x,t,I)\sim\mathcal{D}} \left[\mathcal{L}_{\text{CE}}(W_{\text{F}} W_{\text{CBL}} \psi(x,t), I) + \frac{\lambda}{|\text{D}|}\Omega(W_{\text{CBL}})\right], \tag{11}$$

где $\Omega(W_{\text{CBL}})$ соответсвует регуляризатору. Мы используем:

$$\Omega(W_{\text{CBL}}) = \|W_{\text{CBL}}\|_1 \tag{12}$$

Результаты: Concept Matrix Search алгоритм

Input: Batch of image embeddings $\mathrm{I}_{|\mathrm{B}|}$, labels, all classes C and concepts D embeddings.
Build $\mathcal{V} \in \mathbb{R}^{|\mathrm{B}| \times |\mathrm{D}|}$, $\mathcal{T} \in \mathbb{R}^{|\mathrm{C}| \times |\mathrm{D}|}$ matrices, store $\mathcal{T}$.
for $\mathrm{k} = 0, 1, 2, \ldots, |\mathrm{B}| - 1$ **do**
    for $\mathrm{m} = 0, 1, 2, \ldots, |\mathrm{C}| - 1$ **do**
        Compute and store $\cos(\mathcal{V}^{\top}_{\mathrm{k},.}, \mathcal{T}^{\top}_{\mathrm{m},.})$
    end for
    Find $\mathrm{m}_{\max} = \max_{\mathrm{m}} \cos(\mathcal{V}^{\top}_{\mathrm{k},.}, \mathcal{T}^{\top}_{\mathrm{m},.})$
    if $\mathrm{label}(\mathrm{k}) = \mathrm{m}_{\max}$ **then**
        the hypothesis has been proven, increase Accuracy
    else
        the hypothesis has been disproved
    end if
end for
return Mean accuracy

Сравнение наших методов с предшествующими

Таблица: Сравнение перформанса Bottleneck моделей на основных датасетах. Мы наблюдаем превосходство Sparse-CBM над другими архитектурами на CIFAR10, CIFAR100 и CUB200 датасетах.

Model	CIFAR10	CIFAR100	ImageNet	CUB200	Places
Sparse-CBM	**91.17%**	**74.88%**	71.61%	**80.02%**	41.34%
ℓ_1-CBM	85.11%	73.24%	71.02%	74.91%	40.87%
Contr-CBM	84.75%	68.46%	70.22%	67.04%	40.22%
[2]	86.40%	65.13%	**71.95%**	74.31%	**43.68%**
[3]	83.34%	57.20%	62.57%	63.92%	39.66%
[4]	87.90%	69.10%	70.40%	71.80%	39.43%
Probing	96.12%	80.03%	83.90%	79.29%	48.33%

Сравнение Concept Matrix Search алгоритма с предшествующими методами

Таблица: Сравнение CMS и "DescriptionCLS"[5] на основных датасетах.

Method	CIFAR10	CIFAR100	ImageNet	CUB200	Places
CMS	**85.03%**	62.95%	**77.82%**	**65.17%**	39.43%
[5]	81.61%	**68.32%**	75.00%	63.46%	40.55%
Zero-shot	81.79%	52.84%	76.20%	62.63%	**41.12%**

Визуализация пространства эмбеддингов

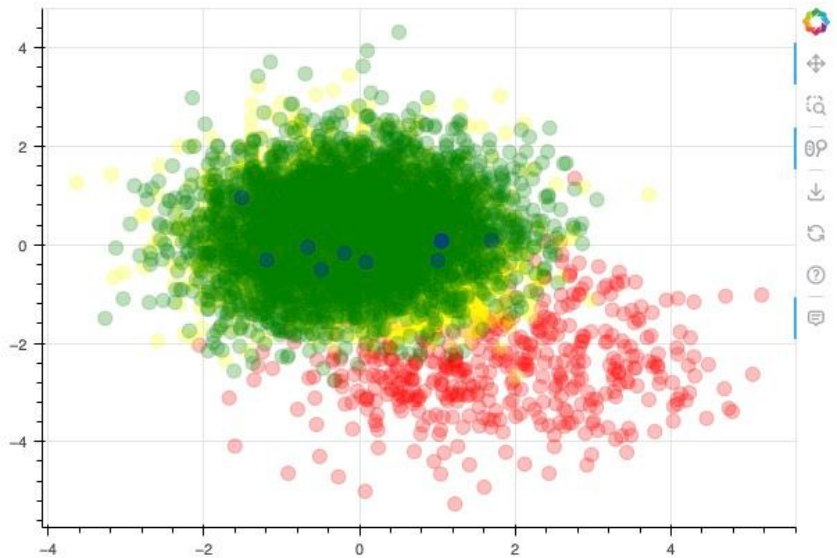


Рис.: CIFAR10 t-SNE. Зелёные точки – проекция эмбеддингов концептов, синие – проекция эмбеддингов классов, красные – картинки, и желтые – случайные слова.

Визуализация CBM

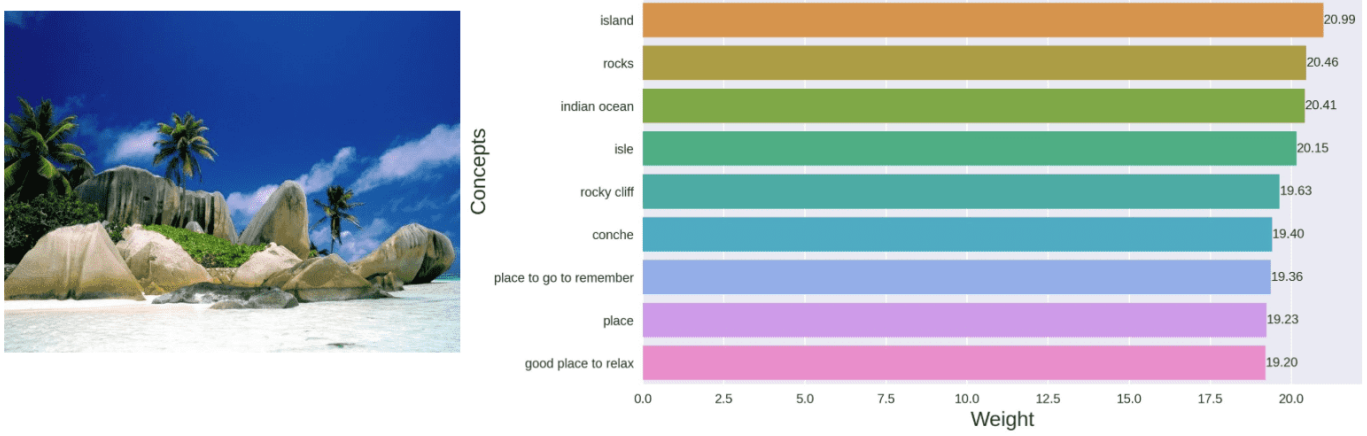


Рис.: Концепты, извлекаемые с помощью CLIP.

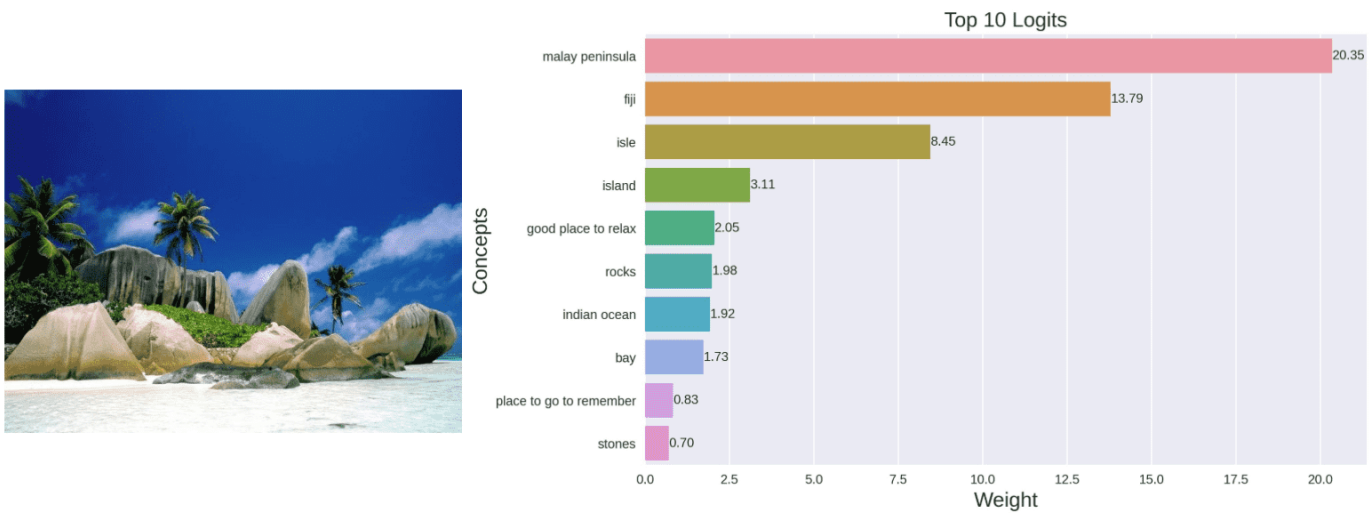


Рис.: Концепты, извлекаемые с помощью Sparse-CBM.

Визуализация CBM

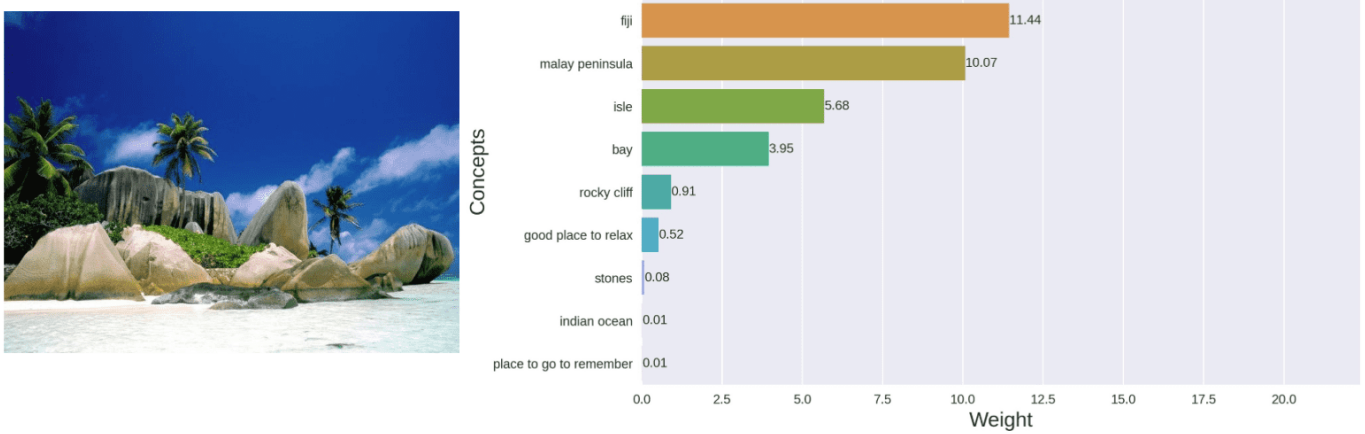


Рис.: Концепты, извлекаемые с помощью ℓ_1-CBM.

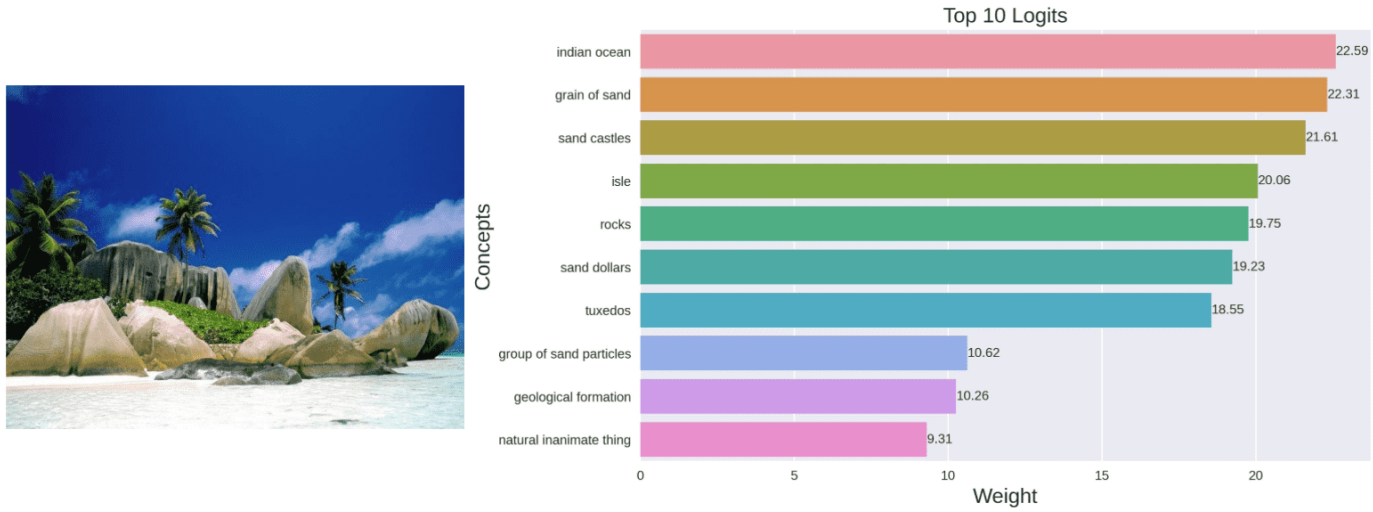


Рис.: Концепты, извлекаемые с помощью Contrastive-CBM.

Ссылки

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
[2] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In The Eleventh International Conference on Learning Representations, 2023.
[3] Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models, 2023.
[4] Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for in- terpretable image classification, 2023.
[5] Menon, S. and Vondrick, C. Visual classification via description from large language models. ICLR, 2023.