

# Анализ смещения распределений в контрастивном обучении

Лидия Троешестова   Роман Исаченко

Московский физико-технический институт

19 мая 2023 г.

## Цель

Исследовать влияние смещения распределения  $p_x^+$  в задаче построения представлений без учителя.

## Проблема

Наличие смещения в распределениях классов приводит к некорректным представлениям объектов.

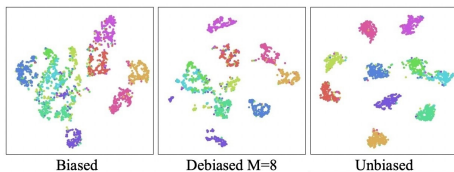
## Идея

Учесть смещения распределений классов путем построения несмещенной функции потерь.

$$L_{\text{Unbiased}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p, \mathbf{x}^+ \sim p_x^+, \\ \mathbf{x}^- \sim p_x^-}} \left[ -\log \frac{\exp(\text{sim}_f(\mathbf{x}, \mathbf{x}^+))}{\exp(\text{sim}_f(\mathbf{x}, \mathbf{x}^+) + \sum_{i=1}^N \exp(\text{sim}_f(\mathbf{x}, \mathbf{x}^-))} \right]$$

# Смещение распределений при построении представлений

- $p_x^+(\mathbf{x}')$  — вероятность взять  $\mathbf{x}'$  как позитивный объект для  $\mathbf{x}$ .
- $p_x^-(\mathbf{x}')$  — вероятность взять  $\mathbf{x}'$  как негативный объект для  $\mathbf{x}$ .
- $\tau^+$  — вероятность 1 класса;
- $\tau^- = 1 - \tau^+$  — вероятность любого другого класса
- $p(\mathbf{x}') = \tau^+ p_x^+(\mathbf{x}') + \tau^- p_x^-(\mathbf{x}')$



Найти  $L_{\text{DebiasedPos}}^N$ , минимизирующее  $\lim_{N \rightarrow \infty} |L_{\text{DebiasedPos}}^N(f) - L_{\text{Unbiased}}^N(f)|$ .

$$L_{\text{Unbiased}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p, \mathbf{x}^+ \sim p_x^+, \\ \mathbf{x}^- \sim p_x^-}} \left[ -\log \frac{\exp(\text{sim}_f(\mathbf{x}, \mathbf{x}^+))}{\exp(\text{sim}_f(\mathbf{x}, \mathbf{x}^+) + \sum_{i=1}^N \exp(\text{sim}_f(\mathbf{x}, \mathbf{x}^-))} \right]$$



Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, Stefanie Jegelka (2020)

Debiased Contrastive Learning



Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, Dilip Krishnan (2021)

Supervised Contrastive Learning



CTing Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton (2020)  
A Simple Framework for Contrastive Learning of Visual Representations



Sohn, Kihyuk (2016)

Improved Deep Metric Learning with Multi-class N-pair Loss Objective



Florian Schroff, Dmitry Kalenichenko, James Philbin (2015)

FaceNet: A Unified Embedding for Face Recognition and Clustering

Обозначим  $h(\mathbf{x}, \tilde{\mathbf{x}}) := e^{f(\mathbf{x})^T f(\tilde{\mathbf{x}})}$ .

## Лемма

При  $N \rightarrow \infty$ :

$$L_{Unbiased}^N(f) \longrightarrow \mathbb{E}_{\substack{\mathbf{x} \sim p \\ \mathbf{x}^- \sim p_x^-}} \left[ -\log \frac{R}{R + N \mathbb{E}_{\mathbf{x}^- \sim p_x^-} h(\mathbf{x}, \mathbf{x}^-)} \right],$$

где

$$R = \frac{1}{\tau^+} (\mathbb{E}_{\mathbf{x}' \sim p} h(\mathbf{x}, \mathbf{x}') - \tau^- \mathbb{E}_{\mathbf{x}^- \sim p_x^-} h(\mathbf{x}, \mathbf{x}^-)).$$

$$\tilde{L}_{DebiasedPos}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p \\ \mathbf{x}^- \sim p_x^-}} \left[ -\log \frac{\mathbb{E}_{\mathbf{x}' \sim p} h(\mathbf{x}, \mathbf{x}') - \tau^- \mathbb{E}_{\mathbf{x}^- \sim p_x^-} h(\mathbf{x}, \mathbf{x}_i^-)}{\mathbb{E}_{\mathbf{x}' \sim p} h(\mathbf{x}, \mathbf{x}') + (N\tau^+ - \tau^-) \mathbb{E}_{\mathbf{x}^- \sim p_x^-} h(\mathbf{x}, \mathbf{x}_i^-)} \right]$$

Оценим неизвестные матожидания эмпирически:

$$P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \mathbf{v}) = \frac{1}{N+2} \left( \sum_{i=1}^N h(\mathbf{x}, \mathbf{u}_i) + h(\mathbf{x}, \mathbf{v}) + h(\mathbf{x}, \mathbf{x}) \right); P_{\text{emp}}^-(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N) = \frac{1}{N} \sum_{i=1}^N h(\mathbf{x}, \mathbf{u}_i).$$

$$\tilde{L}_{\text{DebiasedPos}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p \\ \mathbf{x}^- \sim p_x^-}} \left[ -\log \frac{\mathbb{E}_{\mathbf{x}' \sim p} h(\mathbf{x}, \mathbf{x}') - \tau^- \mathbb{E}_{\mathbf{x}^- \sim p_x^-} h(\mathbf{x}, \mathbf{x}_i^-)}{\mathbb{E}_{\mathbf{x}' \sim p} h(\mathbf{x}, \mathbf{x}') + (N\tau^+ - \tau^-) \mathbb{E}_{\mathbf{x}^- \sim p_x^-} h(\mathbf{x}, \mathbf{x}_i^-)} \right]$$

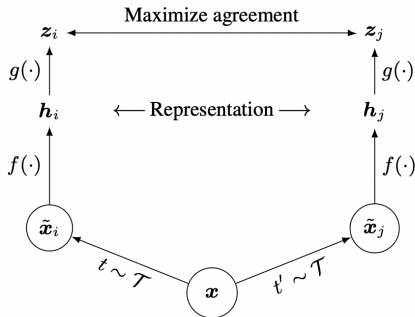
Финальная оценка:

$$L_{\text{DebiasedPos}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p \\ \{\mathbf{u}_i\}_{i=1}^N \sim p_x^- \\ \mathbf{v} \sim p_x^+}} \left[ -\log \frac{P_{\text{emp}} - \tau^- P_{\text{emp}}^-}{P_{\text{emp}} + (N\tau^+ - \tau^-) P_{\text{emp}}^-} \right].$$

## Теорема

Для произвольного представления  $f$  и произвольного  $\delta > 0$  существует достаточно большое  $N$ , что

$$|\tilde{L}_{\text{DebiasedPos}}^N(f) - L_{\text{DebiasedPos}}^N(f)| \leq \left[ \left(1 + \frac{\tau^-}{\tau^+} + \delta\right) \sqrt{\frac{\pi}{2N}} + \left(1 + \frac{1}{\tau^+}\right) \sqrt{\frac{\pi}{2N+2}} \right] e^{3/2}$$



- $\mathcal{T}$  – семейство аугментаций (color distortion, Gaussian blur)
- Семплируются 2 аугментации  $t, t' \sim \mathcal{T}$ , применяются к каждому объекту.
- Обучаем сеть-энкодер  $f(\cdot)$  и MLP сеть-проекцию  $g(\cdot)$ , максимизируя соответствие представлений.

## Цель эксперимента

Сравнить качество представлений при использовании 3 функций потерь: Contrastive, DebaisedNeg и DebaisedPos.

## Обучение

- Обучаем SimCLR с энкодером ResNet-18 и оптимизатором Adam, 50 эпох с размером батча 512.
- Фиксируем выученные представления изображений из CIFAR10 на обучающей выборке.

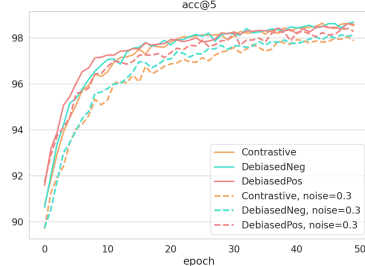
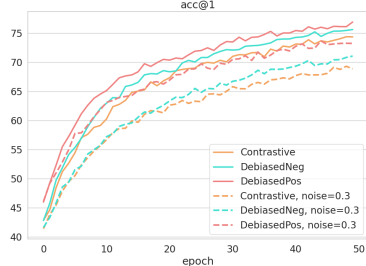
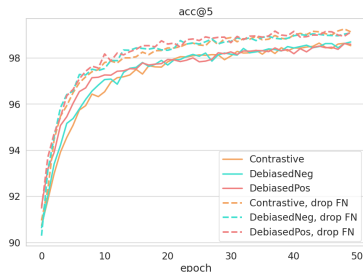
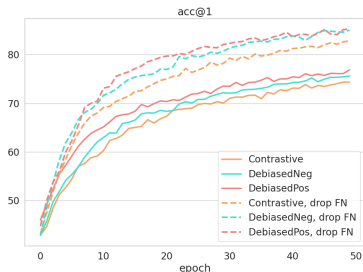
## Валидация

На тестовой выборке классифицируем объект, применяя top-K к банку представлений, считаем top-1 и top-5 accuracy.



# Эксперимент с удалением FN, эксперимент с добавлением FP

Функция потерь DebiasedPos устойчива к шуму и имеет превосходство на ранних эпохах.



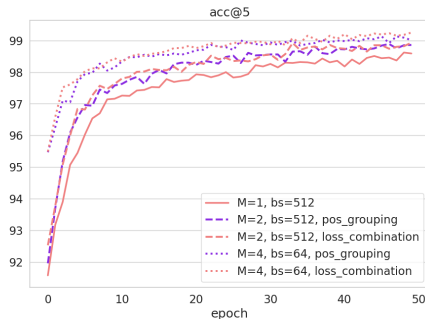
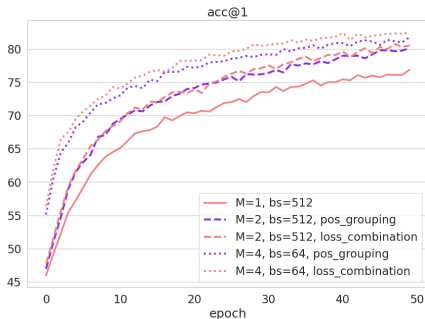
# Увеличение кол-ва позитивных объектов $M$

Способы агрегации по  $M$ :

- 1 *pos-grouping*: внутри оценки  $P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{v}_j\}_{j=1}^M)$  вместо  $h(\mathbf{x}, \mathbf{v})$  используем

$$\frac{1}{M} \sum_{j=1}^M h(\mathbf{x}, \mathbf{v}_j).$$

- 2 *loss-combination*: для каждой пары позитивных объектов считаем  $L_{\text{DebiasedPos}}^N(f)$  и берем среднее по всем значениям функции потерь.

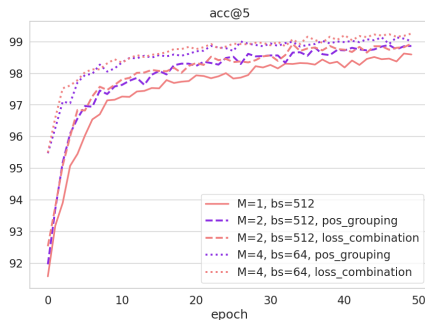
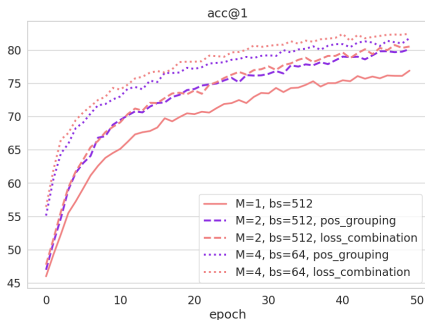


При увеличении  $M$  точность значительно возрастает, причем *loss-combination* лучше, чем *pos-grouping*.

# Увеличение кол-ва позитивных объектов $M$

Method	M	bs	Training time (min)	acc@1	acc@5
-	1	512	122.85	76.76	98.61
pos-grouping	2	512	<b>181.87</b>	80.1	98.86
loss-combination	2	512	192.82	<b>80.5</b>	<b>98.93</b>
pos-grouping	4	64	<b>278.88</b>	81.75	99.01
loss-combination	4	64	308.25	<b>82.44</b>	<b>99.25</b>

Есть trade-off между производительностью и точностью.



- При верности предположения о правильности негативных объектов функция потерь `DebiasedPos` работает корректно.
- `DebiasedPos` более устойчив к шумным датасетам: когда увеличена доля ошибок I рода, `DebiasedPos` имеет большое преимущество в точности.
- При увеличенном кол-ве позитивных объектов точность для всех рассмотренных функций потерь возрастает.
- Способ агрегации *pos-grouping* менее затратный, а *loss-combination* более точный.