
АНАЛИЗ СМЕЩЕНИЯ РАСПРЕДЕЛЕНИЙ ПРИ ИСПОЛЬЗОВАНИИ СРАВНИТЕЛЬНОГО ПОДХОДА В ОБУЧЕНИИ ПРЕДСТАВЛЕНИЙ ДАННЫХ

Мария Никитина
nikitina.mariia@phystech.edu

Роман Исаченко
roman.isachenko@phystech.edu

23 марта 2024 г.

АННОТАЦИЯ

В последнее время contrastive learning вновь приобретает популярность. Данный метод заключается в сравнении положительных (похожих) и отрицательных (непохожих) пар из выборки для получения представления без меток. Однако наличие ложноотрицательных и ложноположительных пар вследствие шумов и погрешностей разметки приводит к смещению функции потерь, не учитывающей наличия таких элементов. В данной работе анализируются различные способы устранения этих искажений с целью точнее восстановить исходное распределение данных. Основываясь на примере из области обучения с учителем, разрабатывается несмещённая модель contrastive learning, исследуются её свойства. Качество несмещённого представления оценивается в задаче классификации, в задаче Image-Text retrieval, а также в искусственном эксперименте на примере двумерного пространства.

Ключевые слова: Contrastive learning · Representation learning · Self-supervised learning

1 Введение

В последние годы все большую популярность приобретают методы обучения без учителя. Они хороши тем, что не требуют предварительной разметки данных, а, значит, для них можно быстрее и проще найти большие по объёму выборки для обучения.

Self-supervised learning – класс методов машинного обучения, позволяющих модели обучаться на задачах, которые не требуют явной разметки данных. Вместо этого используются некоторые характеристики выборки для создания своей собственной разметки.

В частности, *representation learning* – это процесс отображения необработанных входных данных в пространство векторов-признаков или тензоров с целью найти и

извлечь полезные закономерности, которые помогут в предсказании целевых значений. Термину Representation Learning часто ставят в синонимы второй термин – *Feature learning* – машинное обучение, направленное на выделение нужных признаков из данных

На заре машинного обучения много усилий тратилось на разработку методов преобразования данных и предварительной обработки, а модель использовалась только для принятия поверхностных решений на основе извлеченных признаков. Одной из ключевых составляющих успеха глубокого обучения является способность изучать и извлекать с помощью глубоких слоев некоторые полезные признаки из данных. Увеличение объема доступных вычислений и размеченных наборов данных позволило перейти от использования средств извлечения признаков, разработанных вручную, к средствам извлечения посредством изучения данных. В результате фокус исследований также сместился с разработки признаков на разработку архитектуры. Исследования в области архитектуры глубокого обучения в последние годы резко возросли и вылились в несколько основных принципов и строительных блоков, например, свёрточный слой, рекуррентный слой и attention-слой.

К сожалению, задача поиска хорошего представления может быть сложной. Согласно [10], существует несколько признаков, на которых строится representation learning: локальная гладкость, связность в пространстве и времени для последовательности наблюдений, наличие простых зависимостей между признаками. Одно из подходящих решений – *contrastive learning* – подход при котором обучение происходит не только по принципу близости, но и по принципу различия. В отличие от дискриминативной модели, которая учится моделировать границу принятия решений среди классов, и генеративной модели, которая реконструирует входные выборки, при contrastive learning представление изучается путем сравнения входных выборок. Вместо того чтобы изучать сигнал по отдельным выборкам данных по одному элементу за раз, модель contrastive learning обучается путем сравнения различных выборок. Положим вектор \mathbf{x} некоторого объекта в качестве основного. Тогда вектор схожего объекта назовём \mathbf{x}^+ – позитивный элемент. Он должен быть как можно ближе к \mathbf{x} . А вектор отличного объекта \mathbf{x}^- – как можно дальше, как негативный элемент. Основной сложностью использования такого подхода является правильный подбор негативных примеров.

Существует две меры различия элементов: *triplet loss* [12] и его модификация для N негативных примеров – *Multi-Class N-pair loss* [13], а также *Information Noise-Contrastive Estimation* или *InfoNCE* [14].

Triplet loss был впервые использован в задаче распознавания одного и того же человека в разных позах и под разным углом. Пусть есть элемент \mathbf{x} , называемый якорем. Выбирается один положительный элемент \mathbf{x}^+ и один отрицательный \mathbf{x}^- с предположением, что \mathbf{x}^+ принадлежит к тому же классу, что и \mathbf{x} , а \mathbf{x}^- – к другому. Затем расстояние между \mathbf{x} и \mathbf{x}^+ уменьшается, а расстояние между \mathbf{x} и \mathbf{x}^- увеличивается:

$$\mathcal{L}_{triplet}(\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}) = \sum_{x \in \chi} \max(0, \|f(\mathbf{x}) - f(\mathbf{x}^+)\|_2^2 - \|f(\mathbf{x}) - f(\mathbf{x}^-)\|_2^2 + \varepsilon) \quad (1)$$

Обобщение triplet loss:

$$\mathcal{L}_{N\text{-pair}}(\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}_{i=1}^{N-1}; f) = -\log \frac{\exp(f(\mathbf{x})^T f(\mathbf{x}_i^+))}{\exp(f(\mathbf{x})^T f(\mathbf{x}_i^+)) + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^T f(\mathbf{x}_i^-))} \quad (2)$$

Если брать только один негативный элемент из каждого класса, то получится softmax для мультиклассификации.

NCE – это изначально параметр оценки статистической модели. Идея заключалась в том, чтобы использовать логистическую регрессию, чтобы отделить данных от шума. Пусть \mathbf{x} – целевой элемент: $P(\mathbf{x}|C = 1; \theta) = p_\theta(\mathbf{x})$. Пусть $\tilde{\mathbf{x}} : P(\tilde{\mathbf{x}}|C = 0) = q(\tilde{\mathbf{x}})$ – шумовой элемент. Логистическая регрессия моделирует логиты, поэтому моделируется логит выборки u из целевого распределения данных на фоне распределения шума.

$$l_\theta(\mathbf{u}) = \log \frac{p_\theta(\mathbf{u})}{q(\mathbf{u})} = \log p_\theta(\mathbf{u}) - \log q(\mathbf{u}) \quad (3)$$

После преобразования логитов в вероятности с помощью сигмоиды, можно применить кросс-энтропию:

$$\mathcal{L}_{NCE} = -\frac{1}{N} \sum_{i=1}^N [\log \sigma(l_\theta(\mathbf{x}_i)) + \log(1 - \sigma(l_\theta(\tilde{\mathbf{x}}_i)))] \quad (4)$$

$$\text{где } \sigma(l) = \frac{1}{1 + \exp(-l)} = \frac{p_\theta}{p_\theta + q}$$

Методы contrastive learning хорошо себя показали в формировании визуальных представлений. Например, в [3] представлен фреймворк SimCLR, значительно превосходящий предыдущие self-supervised и semi-supervised методы. Его авторы разработали NT-Xent, в котором используется косинусное расстояние вместо скалярных произведений:

$$\mathcal{L}_{SimCLR}^{(i,j)} = -\log \frac{\exp(\text{sim}(g(\mathbf{h}_i), g(\mathbf{h}_j))/\tau)}{\sum_{i \neq k} \exp(\text{sim}(g(\mathbf{h}_i), g(\mathbf{h}_k))/\tau)},$$

где $\mathbf{h}_i = f(\mathbf{x})$, то есть вместо пространства представлений в лосс входит его проекция.

Во время обучения отсутствие истинных меток часто приводит к необходимости случайного выбора отрицательных экземпляров \mathbf{x}^- из данных обучения. Однако такой подход приводит к смещению, если \mathbf{x}^- на самом деле похож на \mathbf{x} . Это смещение приводит к значительному снижению производительности [7].

В [4] был предложен несмещённый контрастивный лосс. Учитывалось N отрицательных и M положительных элементов из выборки. $L_{\text{DebiasedNeg}}^{N,M}(f)$ – функция потерь, которая корректирует ложноотрицательные экземпляры. Чтобы избавиться от необходимости разметки, элементы берутся из распределения данных, а положительное

распределение создаётся искусственно с помощью аугментаций, которые могут содержать ложноположительные выбросы.

Эта работа направлена на разработку нового алгоритма для несмещенного выбора положительных и отрицательных экземпляров выборки, который будет снижать вероятность получить как ложноположительные, так и ложноотрицательные элементы.

2 Предыдущие исследования

Основная задача данного исследования – создание пространства эмбедингов, восстанавливающего начальное распределение, из которого были порождены данные. Идея учёта смещения в лосс-функции путём введения скрытых классов и вероятностей получить для элементов \mathbf{x} и \mathbf{x}^+ один и тот же класс описана в [4] и [1]. В качестве распределения в скрытом пространстве берётся равномерное в предположении, что темы в лучшем пространстве эмбедингов должны быть распределены таким образом. Рассматривается задача классификации на неразмеченных данных. В таком случае важно подобрать для \mathbf{x} отрицательные пары. Ввиду того, что в качестве положительных пар берутся различные аугментации данного изображения, а в качестве отрицательных – все остальные картинки в батче, получается, что неизбежно появятся ложноотрицательные пары, так как не все изображения в выборке имеют класс, отличный от класса элемента \mathbf{x} . В [4] в лосс-функции вероятность получить отрицательную пару выражается через полную вероятность и вероятность получить положительную пару. Тем самым приближается неизвестное распределение отрицательных элементов в выборке.

В [11] и [16] рассматривается применение contrastive learning в задаче сопоставления картинок и текстов для предобучения. В [11] InfoNCE loss используется не только для сопоставления похожих картинок к картинкам, а текстов к текстам, но и для приближения эмбедингов близких по смыслу картинок и текстов в одном пространстве перед тем, как передать их мультимодальной модели. Однако, в [16], в отличие от [11], также учитывается проблема того, что обычно в изображении хранится больше информации, чем описано в тексте. В [16] применяется локальное выравнивание для максимизации взаимной информации между локальным и глобальным представлениями. Итоговая функция потерь в данном фреймворке состоит из кросс-модального выравнивания, где InfoNCE loss применяется для сравнения картинки с текстом и текста с картинкой:

$$\mathcal{L}_{nce}(I, T_+, \{T_k\}_{k=1}^K) = -\mathbb{E}_{p(I, T)} \left[\log \frac{e^{\text{sim}(I, T_+)/\tau}}{\sum_{k=1}^K e^{\text{sim}(I, T_k)/\tau}} \right], \quad (5)$$

где τ – параметр температуры, T_+ – положительный текстовый семпл, $\{T_k\}_{k=1}^K$ – множество негативных текстовых семплов, $\text{sim}(I, T_k) = f_v(v)^\top \hat{f}_t(\hat{t})$, где f_v и \hat{f}_t – проекции выходов визуального и текстового энкодеров на пространство, в котором применяется лосс-функция. Для выбора отрицательных текстовых элементов используется идея, берущая начало с *Memory Bank* [15], который для уменьшения числа вычислений эмбедингов хранит представления с предыдущих итераций. Эта идея имеет развитие в [5], где используются эмбединги не самой модели, а её версии с моментумом. Таким образом, \hat{t}

– эмбединг, полученный с помощью модели с моментумом, а \hat{f}_t – функция, действующая на пространстве таких представлений. То есть параметры модели, вычисляющей хранимые эмбединги, представляют собой скользящее среднее по эпохам.

Аналогично определяется вторая часть данного лосса, высчитываемая относительно текста, а не изображения:

$$\mathcal{L}_{nce}(T, I_+, \{I_k\}_{k=1}^K) = -\mathbb{E}_{p(I,T)} \left[\log \frac{e^{(\text{sim}(T,I)/\tau)}}{\sum_{k=1}^K e^{(\text{sim}(T,I_k)/\tau)}} \right], \quad (6)$$

$$\mathcal{L}_{cma} = \frac{1}{2} [\mathcal{L}_{nce}(I, T_+, \{T_k\}_{k=1}^K) + \mathcal{L}_{nce}(T, I_+, \{I_k\}_{k=1}^K)] \quad (7)$$

Помимо кросс-модального сравнения, итоговая функция потерь включает в себя внутри-модальное выравнивание, где InfoNCE loss применяется для сравнения картинки с картинкой и текста с текстом. Это используется, так как объекты разной модальности зачастую не могут полностью описать друг друга. В таком случае нужно выравнивание ещё и внутри каждой модальности, чтобы приблизить друг к другу различные аугментации одно и того же элемента данных:

$$\mathcal{L}_{imc} = \frac{1}{2} [\mathcal{L}_{nce}(T, T_+, \{T_k\}_{k=1}^K) + \mathcal{L}_{nce}(I, I_+, \{I_k\}_{k=1}^K)] \quad (8)$$

Для того, чтобы учитывать не только глобальную информацию с изображений и текста, в итоговую функцию потерь в [16] входит локальная максимизация, где с помощью InfoNCE loss часть изображения сравнивается со всем изображением, а часть текста – со всем текстом:

$$\mathcal{L}_{lmi} = \frac{1}{2} \left[\frac{1}{M} \sum_{i=1}^M \mathcal{L}_{nce}(T, T_+^i, \{T_k\}_{k=1}^K) + \frac{1}{N} \sum_{j=1}^N \mathcal{L}_{nce}(I, I_+^j, \{I_k\}_{k=1}^K) \right], \quad (9)$$

где T^i – i -ая часть изображения, I^j – j -ая часть текста.

Чтобы соединить представления картинок и изображений, используется задача бинарной классификации, где в качестве таргета используется результат применения лосса из кросс-модального выравнивания, а в качестве предсказания – выход полносвязного слоя общего энкодера:

$$\mathcal{L}_{itm} = \mathbb{E}_{p(I,T)} H(\varphi(I, T), y^{(I,T)}), \quad (10)$$

где $H(\cdot)$ – кросс-энтропия, $y^{(I,T)}$ – таргет.

Маскированная текстовая модель, в которой слова с вероятностью 15% заменяются на специальный токен [MASK] или случайное слово, используется для того, чтобы снизить влияние шума. Предсказание пропущенного слова основывается на контексте текста и изображения:

$$\mathcal{L}_{mlm} = \mathbb{E}_{p(I, T^{msk})} H(\Phi(I, T^{msk}), y^{T^{msk}}), \quad (11)$$

где $\Phi(I, T^{msk})$ – предсказанная вероятность T^{msk} , $y^{T^{msk}}$ – настоящий токен.

Итоговый лосс в [16]:

$$\mathcal{L} = \mathcal{L}_{cma} + \mathcal{L}_{imc} + \mathcal{L}_{lmi} + \mathcal{L}_{itm} + \mathcal{L}_{mlm} \quad (12)$$

Обоснование применения InfoNCE заключается в том, что он связан с нижней границей взаимной информации, которую тяжело посчитать напрямую. В [14] представлено доказательство этого факта с помощью связи лосса с MINE [2].

3 Постановка задачи

3.1 Несмещённый лосс

Пусть \mathcal{X} – предметное пространство. Contrastive learning работает с семантически близкими парами точек $(\mathbf{x}, \mathbf{x}^+)$, где \mathbf{x} берётся из распределения данных $p(x)$ над \mathcal{X} . Цель состоит в том, чтобы найти функцию эмбединга $f : \mathcal{X} \rightarrow \mathbb{R}^d$. В [4] предполагается, что существует множество дискретных скрытых классов \mathcal{C} и положительные пары $(\mathbf{x}, \mathbf{x}^+)$ имеют один и тот же класс. Распределение по \mathcal{C} обозначается $\rho(c)$, следовательно, совместное распределение $p_{x,c}(\mathbf{x}, c) = p(\mathbf{x}|c)\rho(c)$. Пусть $h : \mathcal{X} \rightarrow \mathcal{C}$ – функция, присваивающая метки классов. Тогда $p_x^+(\mathbf{x}') = p(\mathbf{x}'|h(\mathbf{x}') = h(\mathbf{x}))$ – вероятность сказать, что \mathbf{x}' – положительная пара для \mathbf{x} , и $p_x^-(\mathbf{x}') = p(\mathbf{x}'|h(\mathbf{x}') \neq h(\mathbf{x}))$ – для отрицательной пары. Предполагается, что вероятности классов $\rho(\mathbf{x}) = \tau^+$ однородны и $\tau^- = 1 - \tau^+$ – вероятность обнаружить любой другой класс. Тогда «идеальный» лосс при условии наличия N негативных элементов выглядеть так:

$$L_{\text{Unbiased}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p, \mathbf{x}^+ \sim p_x^+, \\ \mathbf{x}_i^- \sim p_x^-}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \frac{Q}{N} \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \quad (13)$$

В знаменателе авторы [4] поставили множитель $\frac{Q}{N}$ для взвешивания и дальнейшего анализа. При первой оценке и конечном N берётся $Q = N$. Тогда:

$$L_{\text{Unbiased}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p, \mathbf{x}^+ \sim p_x^+, \\ \mathbf{x}_i^- \sim p_x^-}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \quad (14)$$

На практике $p_x^-(\mathbf{x}_i^-)$ неизвестно, поэтому \mathbf{x}_i^- выбирается из немаркированного распределения $p(\mathbf{x})$. С вероятностью τ^+ попадётся ложноотрицательный элемент. Лосс для такого случая уже будет смещённым.

$$L_{\text{Biased}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p, \mathbf{x}^+ \sim p_x^+, \\ \mathbf{x}_i^- \sim p_x}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \quad (15)$$

В [4] доказывается лемма о том, что в пределе $L_{\text{Biased}}^N(f)$ – верхняя грань $L_{\text{Unbiased}}^N(f)$.

Задачей является поиск наиболее близкой к $L_{\text{Unbiased}}^N(f)$ функции потерь при условии наличия M положительных элементов и распределения $p(\mathbf{x})$. Его разложение:

$$p(\mathbf{x}') = \tau^+ p_x^+(\mathbf{x}') + \tau^- p_x^-(\mathbf{x}') \quad (16)$$

Выразим $p_x^-(\mathbf{x}')$:

$$p_x^-(\mathbf{x}') = \frac{p(\mathbf{x}') - \tau^+ p_x^+(\mathbf{x}')}{\tau^-} \quad (17)$$

Подсчёт $p_x^-(\mathbf{x}')$ через $p(\mathbf{x})$ и $p_x^-(\mathbf{x}')$ в итоговой формуле возможен, но очень дорогостоящий. Поэтому предлагается в [4] использовать приближение итоговой функции потерь и использовать уже упомянутый выше $L_{\text{DebiasedNeg}}^{N,M}(f)$:

$$L_{\text{DebiasedNeg}}^{N,M}(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p; \mathbf{x}_+ \sim p_x^+, \\ \{\mathbf{u}_i\}_{i=1}^N \sim p^N, \\ \{\mathbf{v}_j\}_{j=1}^M \sim p_x^{+M}}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+) + Ng(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{v}_j\}_{j=1}^M)}} \right], \quad (18)$$

где эмпирическая оценка $p(\mathbf{x}^-)$:

$$g(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{v}_j\}_{j=1}^M) = \max \left\{ \frac{1}{\tau^-} \left(\frac{1}{N} \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)} - \tau^+ \frac{1}{M} \sum_{j=1}^M e^{f(\mathbf{x})^T f(\mathbf{v}_j)} \right), e^{-1/t} \right\}. \quad (19)$$

$\{\mathbf{u}_i\}_{i=1}^N$ – N сэмплов из $p(\mathbf{x}')$ и $\{\mathbf{v}_j\}_{j=1}^M$ – M сэмплов из $p(\mathbf{x}'^+)$, t – параметр температуры.

В [4] также приводится оценка скорости сходимости $L_{\text{DebiasedNeg}}^{N,M}(f)$ к случаю бесконечного N и $M = 1$:

$$|\tilde{L}_{\text{DebiasedNeg}}^N(f) - L_{\text{DebiasedNeg}}^{N,M}(f)| \leq \frac{e^{3/2}}{\tau^-} \sqrt{\frac{\pi}{2N}} + \frac{e^{3/2}\tau^+}{\tau^-} \sqrt{\frac{\pi}{2M}} \quad (20)$$

Применение данной лосс-функции в SimCLR позволило увеличить точность результата и уменьшить число ложноотрицательных элементов. Поэтому основная задача данной работы – разработка алгоритма, направленного на уменьшение числа ложноположительных сэмплов.

3.2 Ложноположительные элементы

Чрезмерная аугментация в попытке получить как можно большее M приводит к появлению ложноположительных сэмплов и смещению оценки. Оценим распределение положительных элементов в рамках функции потерь.

$$p_x^+(\mathbf{x}') = \frac{p(\mathbf{x}') - \tau^- p_x^-(\mathbf{x}')}{\tau^+} \quad (21)$$

В отличие от пункта выше здесь положительная выборка из $p(\mathbf{x}')$ может включать в себя на самом деле отрицательные сэмплы.

Лемма 1. При $N \rightarrow \infty$:

$$\mathbb{E}_{\substack{\mathbf{x} \sim p \\ \mathbf{x}^+ \sim p_x^+ \\ \{\mathbf{x}_i^-\}_{i=1}^N \sim p_x^-}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \rightarrow \mathbb{E}_{\mathbf{x} \sim p} \left[-\log \frac{R}{R + N \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}} \right], \quad (22)$$

где

$$R = \frac{1}{\tau^+} (\mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x}')^T f(\mathbf{x}')} - \tau^- \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}). \quad (23)$$

Доказательство. Поскольку выражение внутри математического ожидания ограничено, мы можем применить теорему Лебега о мажорируемой сходимости:

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathbb{E} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] &= \mathbb{E} \left[\lim_{N \rightarrow \infty} -\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] = \\ &= \mathbb{E} \left[-\log \frac{\mathbb{E}_{\mathbf{x}^+ \sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{\mathbb{E}_{\mathbf{x}^+ \sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + N \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}} \right] \end{aligned}$$

Используя 21 и линейность матожидания, получим:

$$\mathbb{E}_{\mathbf{x}^+ \sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} = \frac{1}{\tau^+} (\mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x}')^T f(\mathbf{x}')} - \tau^- \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}) = R. \quad \blacksquare$$

Назовём предел в лемме $\tilde{L}_{\text{DebiasedPos}}^N(f)$:

$$\tilde{L}_{\text{DebiasedPos}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p \\ \mathbf{x}^- \sim p_x^-}} \left[-\log \frac{\mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x}')^T f(\mathbf{x}')} - \tau^- \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}}{\mathbb{E}_{\mathbf{x}' \sim p} e^{f(\mathbf{x}')^T f(\mathbf{x}')} + (N\tau^+ - \tau^-) \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \quad (24)$$

Полученная функция потерь при конечном N :

$$L_{\text{DebiasedPos}}^N(f) = \mathbb{E}_{\substack{\mathbf{x} \sim p \\ \{\mathbf{u}_i\}_{i=1}^N \sim p_x^- \\ \mathbf{v} \sim p_x^+}} \left[-\log \frac{P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \mathbf{v}) - \tau^- P_{\text{emp}}^-(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N)}{P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \mathbf{v}) + (N\tau^+ - \tau^-) P_{\text{emp}}^-(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N)} \right], \quad (25)$$

где

$$P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \mathbf{v}) = \frac{1}{N+2} \left(\sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})} \right); \quad (26)$$

$$P_{\text{emp}}^-(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N) = \frac{1}{N} \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)}. \quad (27)$$

В [14] представлено доказательство того факта, что InfoNCE loss максимизирует нижнюю границу взаимной информации.

Теорема 2. $\mathcal{L}_{\text{DebiasedPos}}$ максимизирует нижнюю границу взаимной информации.

Доказательство. Пусть c – контекст, x – элемент выборки. Тогда взаимная информация:

$$I(x, c) = \sum_{x \in X, c \in C} p(x, c) \log \frac{p(x|c)}{p(x)}$$

Так как рассматриваемый лосс – это категориальная кросс-энтропия классификации положительного сэмпла корректно, то оптимальная для него вероятность того, что x_i – положительный элемент:

$$p(x_i - \text{положительный} | X, x) = \frac{p(x_i|c) \prod_{l \neq i} p(x_l)}{\sum_{j=1}^N p(x_j|c) \prod_{l \neq j} p(x_l)} = \frac{\frac{p(x_i|c)}{p(x_i)}}{\sum_{j=1}^{N+1} \frac{p(x_j|c)}{p(x_j)}}$$

Так как лосс-функцию можно представить в виде:

$$\mathcal{L} = -\mathbb{E}_{x \sim p} \left(\log \frac{f(x, c)}{\sum_{x_j \in X} f(x_j, c)} \right),$$

то $\frac{p(x_i|c)}{p(x_i)}$ – оптимальное значение для $f(x_i, c)$. Тогда:

$$\begin{aligned} \mathcal{L}_{\text{DebiasedPos}} &= \mathbb{E}_{\substack{x \sim p \\ x_i^- \sim p_x^-}} \left[-\log \frac{\mathbb{E}_{x \sim p} f(x, c) - \tau^- \mathbb{E}_{x^- \sim p_x^-} f(x_i^-, c)}{\mathbb{E}_{x \sim p} f(x, c) + (N\tau^+ - \tau^-) \mathbb{E}_{x^- \sim p_x^-} f(x_i^-, c)} \right] = \\ &= \mathbb{E}_{\substack{x \sim p \\ x_i^- \sim p_x^-}} \left[\log \frac{\mathbb{E}_{x \sim p} \frac{p(x|c)}{p(x)} + (N\tau^+ - \tau^-) \mathbb{E}_{x^- \sim p_x^-} \frac{p(x_i^-|c)}{p(x)}}{\mathbb{E}_{x \sim p} \frac{p(x|c)}{p(x)} - \tau^- \mathbb{E}_{x^- \sim p_x^-} \frac{p(x_i^-|c)}{p(x)}} \right] \approx \\ &\approx \mathbb{E}_{\substack{x \sim p \\ x_i^- \sim p_x^-}} \left[\log \frac{\frac{1}{N+2} \left(\frac{p(x|c)}{p(x)} + \sum_{i=1}^N \frac{p(x_i^-|c)}{p(x_i^-)} + \frac{p(c|c)}{p(c)} \right) + (N\tau^+ - \tau^-) \frac{1}{N} \sum_{i=1}^N \frac{p(x_i^-|c)}{p(x_i^-)}}{\frac{1}{N+2} \left(\frac{p(x|c)}{p(x)} + \sum_{i=1}^N \frac{p(x_i^-|c)}{p(x_i^-)} + \frac{p(c|c)}{p(c)} \right) - \tau^- \frac{1}{N} \sum_{i=1}^N \frac{p(x_i^-|c)}{p(x_i^-)}}} \right] \approx \end{aligned}$$

$$\begin{aligned}
&\approx \mathbb{E}_{\substack{x \sim p \\ x_i^- \sim p_x^-}} \left[\log \frac{\frac{p(x|c)}{p(x)} + N \mathbb{E}_{x_i^- \sim p_x^-} \frac{p(x_i^-|c)}{p(x_i^-)} + 1 + (N\tau^+ - \tau^-)(N+2) \mathbb{E}_{x_i^- \sim p_x^-} \frac{p(x_i^-|c)}{p(x_i^-)}}{\frac{p(x|c)}{p(x)} + N \mathbb{E}_{x_i^- \sim p_x^-} \frac{p(x_i^-|c)}{p(x_i^-)} + 1 - \tau^-(N+2) \mathbb{E}_{x_i^- \sim p_x^-} \frac{p(x_i^-|c)}{p(x_i^-)}} \right] = \\
&= \mathbb{E}_{x \sim p} \left[\log \frac{\frac{p(x|c)}{p(x)} + N + 1 + N^2\tau^+ + 2N\tau^+ - N\tau^- - 2\tau^-}{\frac{p(x|c)}{p(x)} + N + 1 - (N+2)\tau^-} \right] = \\
&= \mathbb{E}_{x \sim p} \log \left[1 + \frac{N(N+2)\tau^+}{\frac{p(x|c)}{p(x)} + N\tau^+ + 1 - 2\tau^-} \right] = \mathbb{E}_{x \sim p} \log \left[1 + \frac{N(N+2)\tau^+ \frac{p(x)}{p(x|c)}}{1 + (N\tau^+ + 1 - 2\tau^-) \frac{p(x)}{p(x|c)}} \right] \geq \\
&\geq \mathbb{E}_{x \sim p} \log \left[\frac{N(N+2)\tau^+ \frac{p(x)}{p(x|c)}}{1 + (N\tau^+ + 1 - 2\tau^-) \frac{p(x)}{p(x|c)}} \right] = -I(x, c) + \mathbb{E}_{x \sim p} \log \left[\frac{N(N+2)\tau^+}{1 + (N\tau^+ + 1 - 2\tau^-) \frac{p(x)}{p(x|c)}} \right] = \\
&= -I(x, c) + \mathbb{E}_{x \sim p} \log \left[\frac{(N+2)\tau^+}{\frac{1}{N} + \left(\tau^+ + \frac{1}{N} - \frac{2\tau^-}{N} \right) \frac{p(x)}{p(x|c)}} \right]
\end{aligned}$$

При $N \rightarrow \infty$ получим:

$$\begin{aligned}
&-I(x, c) + \mathbb{E}_{x \sim p} \log \left[\frac{(N+2)\tau^+}{\frac{1}{N} + \left(\tau^+ + \frac{1}{N} - \frac{2\tau^-}{N} \right) \frac{p(x)}{p(x|c)}} \right] \geq -I(x, c) + \mathbb{E}_{x \sim p} \log \left[\frac{(N+2)\tau^+}{\tau^+ \frac{p(x)}{p(x|c)}} \right] \geq \\
&\geq -I(x, c) + \mathbb{E}_{x \sim p} \log \left((N+2) \frac{p(x|c)}{p(x)} \right)
\end{aligned}$$

Следовательно,

$$I(x, c) \geq \mathbb{E}_{x \sim p} \log \left((N+2) \frac{p(x|c)}{p(x)} \right) - \mathcal{L}_{\text{DebiasedPos}}$$

■

4 Вычислительный эксперимент

4.1 Классификация

В качестве первого вычислительного эксперимента проводится сравнение $\mathcal{L}_{N\text{-pair}}(2)$ и $\mathcal{L}_{\text{DebiasedPos}}^N(f)$ (25) в задаче классификации. В качестве датасета используется CIFAR10

[9], состоящий из 60000 цветных изображений размером 32x32 для классификации. Бейзлайн-модель: SimCLR [4] с энкодером ResNet18 [6] и N-pair loss (2), оптимизатор – Adam [8], learning rate 0.001, размер батча 512. Все модели тренируются 50 эпох и оцениваются линейным классификатором после получения эмбединга.

Сравнение результатов работы можно увидеть на рис.???. В качестве метрики берётся топ-1 и топ-5 accuracy:

$$Acc_1 = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Acc_k = \frac{1}{n} \sum_{i=1}^n [y_i \in \hat{y}_i^k]$$

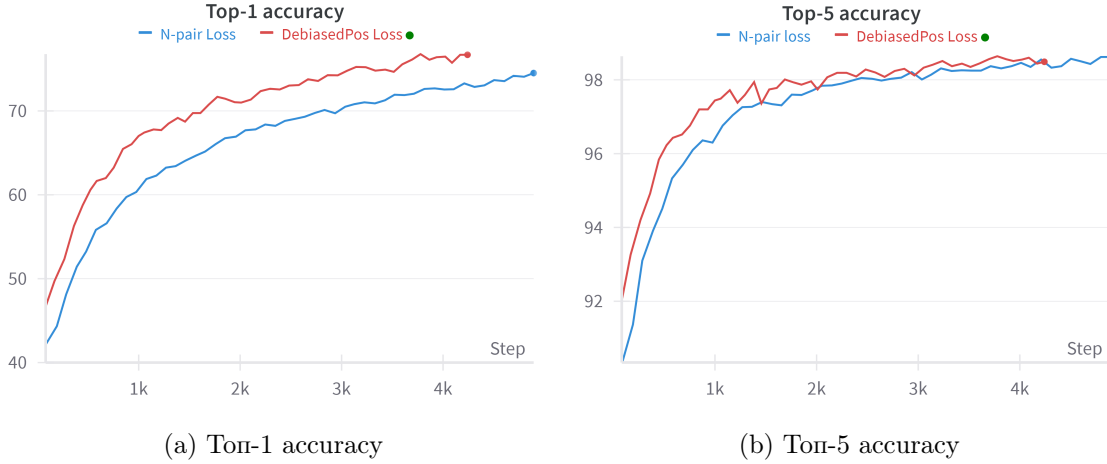


Рис. 1: Метрики классификации с использованием InfoNCE loss и DebiasedPos loss.

4.2 Анализ ошибки

5 Заключение

Список литературы

- [1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, 2019.
- [2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2021.
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

- [4] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debaised contrastive learning, 2020.
- [5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [7] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26449–26461. Curran Associates, Inc., 2021.
- [8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [9] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.
- [10] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.
- [11] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.
- [12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015.
- [13] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Neural Information Processing Systems*, 2016.
- [14] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.
- [15] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018.
- [16] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. 2022.