Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Никитина Мария Александровна

Анализ смещения распределений при использовании сравнительного подхода в обучении представлений данных

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
к.ф.-м.н. Р. В. Исаченко

Москва — 2024

Аннотация

Среди подходов обучения без учителя contrastive learning в последнее время снова приобретает популярность. Данный метод заключается в сравнении пар из выборки для получения пространства представления, в котором похожие элементы как положительные будут находится близко, а отличающиеся – далеко, как отрицательные. 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, разрабатывется 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

Содержание

1 Введение

Методы обучения без учителя становятся достаточно популярными в последнее время. Они хороши тем, что не требуют предварительной разметки данных, а, значит, для них можно проще и быстрее 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. Часто такие методы используются для предварительной обрабоки датасета перед обучением с учителем.

В частности, *representation learning* – это процесс отображения необработанных входных данных в пространство векоторов-признаков или тензоров с целью найти и извлечь полезные закономерности, которые помогут в предсказании целевых значений.

При создании первых моделей 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. Поэтому одной из ключевых составляющих успеха глубокого обучения является способность изучать и извлекать некоторые полезные признаки из данных. Увеличение объема доступных вычислений и размеченных датасетов позволило перейти от использования средств извлечения признаков, разработанных вручную, к автоматическому выделению особенностей. В результате исследования также сместились с разработки признаков на разработку архитектуры модели, которая самостоятельно может создавать представление данных.

К сожалению, задача поиска хорошего представления может быть сложной. Согласно [11], существует несколько принципов, на которых строится representation learning: выразительность и способность кодировать экспоненциальное число конфигураций в отличие от методов наподобие one-hot, абстрактность и отсутствие сильной зависимости от локальных небольших изменений, непротиворечивость и обеспечение наглядности представления. 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 пред-

ставление изучается путем сравнения пар элементов. 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 $\mathbf{x}$ в пространстве эмбеддингов. А вектор отличного объекта $\mathbf{x}^-$ – как можно дальше, как негативный элемент. Основной сложностью использования такого подхода является правильный подбор негативных примеров на неразмеченных данных.

Существует две меры различия элементов: *triplet loss* [14] и его модификация для N негативных примеров – *Multi-Class N-pair loss* [15], а также *Information Noise-Contrastive Estimation* или *InfoNCE* [16] как обобщение N-pair loss.

Triplet loss был впервые использован в задаче распознавания одного и того же человека в разных позах и под разным углом. Пусть есть элемент $\mathbf{x}$. 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:

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

Ввиду того, что обычно в качестве положительной пары берётся аугментация того же изображения, а в качестве отрицательной – весь остальной батч, существует обобщение triplet loss на N негативных элементов:

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

Если вместо $\exp(f(\mathbf{x})^T f(\mathbf{c}))$ брать любую функцию, аппроксимирующую $\frac{p(\mathbf{x}|\mathbf{c})}{p(\mathbf{x})}$, получится InfoNCE loss:

$$\mathcal{L}_{InfoNCE} = -\mathbb{E}\left[\log \frac{f(\mathbf{x}, \mathbf{c})}{\sum_{\mathbf{x}' \in X} f(\mathbf{x}', \mathbf{c})}\right]$$

Методы contrastive learning хорошо себя показали в формировании визуальных представлений. Например, в [4] 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:

$$\mathcal{L}_{SimCLR}^{(i,j)} = -\log \frac{\exp(\text{sim}(g(\mathbf{h}_i), g(\mathbf{h}_j)/\tau)}{\sum_{i \neq k}^{2N} \exp(\text{sim}(g(\mathbf{h}_i), g(\mathbf{h}_k)/\tau)},$$

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 качества[8].

В [5] предложен несмещённый контрастивный лосс. Учитывается 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 снижения вероятности смещения при появлении ложноположительных и ложноотрицательные элементов, а также оценку качества восстановления исходного распределения данных.

2 Предыдущие исследования

Основная задача данного исследования – создание пространства эмбеддинов, 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 [5] и [2]. В качестве распределения в скрытом пространстве берётся равномерное в предположении, что темы в лучшем пространстве эмбеддингов 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 [5] 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 [12] и [18] рассматривается применение contrastive learning в задаче сопоставления картинок и текстов для предобучения. В [12] InfoNCE loss используется не только для сопоставления похожих картинок к картинкам, а текстов к текстам, но и для приближения эмбеддингов близких по смыслу картинок и текстов в одном пространстве перед тем, как передать их мультимодальной модели. Однако, в [18], в отличие от [12], также учитывается проблема того, что обычно в изображении хранится больше информации, чем описано в тексте. В [18] 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:

$$\mathcal{L}_{nce}(I, T_+, \{T_k\}_{k=1}^K) = -\mathbb{E}_{p(I,T)}\left[\log \frac{e^{(\text{sim}(I,T_+)/\tau)}}{\sum_{k=1}^K e^{(\text{sim}(I,T_k)/\tau)}}\right], \tag{3}$$

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 энко-

деров на пространство, в котором применяется лосс-функция. Для выбора отрицательных текстовых элементов используется идея, берущая начало с *Memory Bank* [17], который для уменьшения числа вычислений эмбеддингов хранит представления с предыдущих итераций. Эта идея имеет развитие в [6], где используются эмбеддинги не самой модели, а её версии с моментумом. Таким образом, $\hat{t}$ – эмбеддинг, полученный с помощью модели с моментумом, а $\hat{f}_t$ – функция, действующая на пространстве таких представлений. То есть параметры модели, вычисляющей хранимые эмбеддинги, представляют собой скользящее среднее по эпохам.

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

$$\mathcal{L}_{nce}(T, I_+, \{I_k\}_{k=1}^K) = -\mathbb{E}_{p(I,T)}\left[\log \frac{e^{(\mathrm{sim}(T,I)/\tau)}}{\sum_{k=1}^K e^{(\mathrm{sim}(T,I_k)/\tau)}}\right], \tag{4}$$

$$\mathcal{L}_{cma} = \frac{1}{2}[\mathcal{L}_{nce}(I, T_+, \{T_k\}_{k=1}^K) + \mathcal{L}_{nce}(T, I_+, \{I_k\}_{k=1}^K)] \tag{5}$$

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:

$$\mathcal{L}_{imc} = \frac{1}{2}[\mathcal{L}_{nce}(T, T_+, \{T_k\}_{k=1}^K) + \mathcal{L}_{nce}(I, I_+, \{I_k\}_{k=1}^K)] \tag{6}$$

Для того, чтобы учитывать не только глобальную информацию с изображений и текста, в итоговую функцию потерь в [18] входит локальная максимизация, где с помощью InfoNCE loss часть изображения сравнивается со всем изображением, а часть текста – со всем текстом:

$$\mathcal{L}_{lmi} = \frac{1}{2}\left[\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{nce}(T, T_+^i, \{T_k\}_{k=1}^K) + \frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_{nce}(I, I_+^j, \{I_k\}_{k=1}^K)\right], \tag{7}$$

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:

$$\mathcal{L}_{itm} = \mathbb{E}_{p(I,T)} H(\varphi(I,T), y^{(I,T)}), \tag{8}$$

где $H(;)$ – 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:

$$\mathcal{L}_{mlm} = \mathbb{E}_{p(I,T^{msk})} H(\Phi(I,T^{msk}), y^{T^{msk}}), \tag{9}$$

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

Итоговый лосс в [18]:

$$\mathcal{L} = \mathcal{L}_{cma} + \mathcal{L}_{imc} + \mathcal{L}_{lmi} + \mathcal{L}_{itm} + \mathcal{L}_{mlm} \tag{10}$$

Обоснование применения InfoNCE заключается в том, что он связан с нижней границей взаимной информации, которую тяжело посчитать напрямую. В [16] представлено доказательство этого факта с помощью связи лосса с MINE [3].

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 эмбеддинга $f : \mathcal{X} \to \mathbb{R}^d$. В [5] 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 $h : \mathcal{X} \to \mathcal{C}$ – 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 $\tau^- = 1 - \tau^+$ – ворятность обнаружить любой другой класс. Тогда «идеальный» лосс при условии наличия N негативных элементов выглядеть так:

$$\mathcal{L}^N_{\text{Unbiased}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p, \mathbf{x}^+\sim p_x^+, \\ \mathbf{x}_i^-\sim p_x^-}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \tag{11}$$

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.

$$\mathcal{L}^N_{\text{Biased}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p, \mathbf{x}^+\sim p_x^+, \\ \mathbf{x}_i^-\sim p_x}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \right] \tag{12}$$

В [5] доказывается лемма о том, что в пределе при $N \to \infty$ лосс функция $L^N_{\text{Biased}}(f)$ является верхней гранью $L^N_{\text{Unbiased}}(f)$.

Распределение $p(\mathbf{x})$ при условии наличия M положительных элементов можно разложить:

$$p(\mathbf{x}') = \tau^+ p_x^+(\mathbf{x}') + \tau^- p_x^-(\mathbf{x}') \tag{13}$$

Выразим $p_x^-(\mathbf{x}')$:

$$p_x^-(\mathbf{x}') = \frac{p(\mathbf{x}') - \tau^+ p_x^+(\mathbf{x}')}{\tau^-} \tag{14}$$

Подсчёт $p_x^-(\mathbf{x}')$ через $p(\mathbf{x})$ и $p_x^+(\mathbf{x}')$ в итоговой формуле очень дорогостоящий. Поэтому в [5] предлагается использовать приближение итоговой функции потерь:

$$\mathcal{L}_{\text{Neg}}^{N,M}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p;\mathbf{x}_+\sim p_x^+,\\ \{\mathbf{u}_i\}_{i=1}^N\sim p^N\\ \{\mathbf{v}_j\}_{j=1}^M\sim {p_x^+}^M}} \left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + N g\big(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{v}_j\}_{j=1}^M\big)} \right], \tag{15}$$

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

$$g\big(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^N, \{\mathbf{v}_j\}_{j=1}^M\big) = \max\left\{ \frac{1}{\tau^-}\left(\frac{1}{N}\sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{u}_i)} - \tau^+ \frac{1}{M}\sum_{j=1}^M e^{f(\mathbf{x})^T f(\mathbf{v}_j)} \right), e^{-1/t} \right\}. \tag{16}$$

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

Применение данной лосс-функции в модели SimCLR позволило смоделировать смещение при наличии в выборке ложноотрицательных элементов и увеличить точность результата. В данной работе будет произведено её сравнение с лосс-функцией, в которой вместо $p(\mathbf{x}^-)$ будет выражен $p(\mathbf{x}^+)$.

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.

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

Лемма 1. *При $N \to \infty$:*

$$\mathcal{L}^N_{Unbiased}(f) = \mathbb{E}_{\substack{\boldsymbol{x}\sim p \\ \boldsymbol{x}^+\sim p^+_x \\ \{\boldsymbol{x}^-_i\}^N_{i=1}\sim {p^-_x}^N}} \left[-\log \frac{e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^+)}}{e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^+)} + \sum_{i=1}^N e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^-_i)}} \right] \longrightarrow \\ \longrightarrow \mathbb{E}_{\substack{\boldsymbol{x}\sim p \\ \boldsymbol{x}^-\sim p^-_x}} \left[-\log \frac{R}{R + N\mathbb{E}_{\boldsymbol{x}^-\sim p^-_x} e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^-)}} \right], \tag{18}$$

где

$$R = \frac{1}{\tau^+}\left(\mathbb{E}_{\boldsymbol{x}'\sim p} e^{f(\boldsymbol{x})^T f(\boldsymbol{x}')} - \tau^- \mathbb{E}_{\boldsymbol{x}^-\sim p^-_x} e^{f(\boldsymbol{x})^T f(\boldsymbol{x}^-)}\right). \tag{19}$$

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

$$\begin{aligned} &\lim_{N\to\infty} \mathbb{E}\left[-\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}^-_i)}} \right] = \\ &= \mathbb{E}\left[\lim_{N\to\infty} -\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}^-_i)}} \right] = \\ &= \mathbb{E}\left[-\log \frac{\mathbb{E}_{\mathbf{x}^+\sim p^+_x} e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{\mathbb{E}_{\mathbf{x}^+\sim p^+_x} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + N\mathbb{E}_{\mathbf{x}^-\sim p^-_x} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}} \right] \end{aligned} \tag{20}$$

Используя 17 и линейность матожидания, получим:

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

■

3.3 Функция потерь $\mathcal{L}_{\mathbf{Pos}}$

Назовём предел в лемме $\tilde{\mathcal{L}}^N_{\mathrm{Pos}}(f)$:

$$\tilde{\mathcal{L}}^N_{\mathrm{Pos}}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p \\ \mathbf{x}^-\sim p^-_x}} \left[-\log \frac{\mathbb{E}_{\mathbf{x}'\sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')} - \tau^- \mathbb{E}_{\mathbf{x}^-\sim p^-_x} e^{f(\mathbf{x})^T f(\mathbf{x}^-_i)}}{\mathbb{E}_{\mathbf{x}'\sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')} + \left(N\tau^+ - \tau^-\right)\mathbb{E}_{\mathbf{x}^-\sim p^-_x} e^{f(\mathbf{x})^T f(\mathbf{x}^-_i)}} \right] \tag{21}$$

Полученная функция потерь при конечном N:

$$\mathcal{L}_{\text{Pos}}^{N}(f) = \mathbb{E}_{\substack{\mathbf{x}\sim p \\ \{\mathbf{u}_i\}_{i=1}^{N}\sim p_x^{-N} \\ \mathbf{v}\sim p_x^{+}}} \left[-\log \frac{P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^{N}, \mathbf{v}) - \tau^{-} P_{\text{emp}}^{-}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^{N})}{P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^{N}, \mathbf{v}) + \left(N\tau^{+} - \tau^{-}\right) P_{\text{emp}}^{-}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^{N})} \right], \tag{22}$$

где

$$P_{\text{emp}}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^{N}, \mathbf{v}) = \frac{1}{N+2} \left(\sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)} + e^{f(\mathbf{x})^T f(\mathbf{v})} + e^{f(\mathbf{x})^T f(\mathbf{x})} \right); \tag{23}$$

$$P_{\text{emp}}^{-}(\mathbf{x}, \{\mathbf{u}_i\}_{i=1}^{N}) = \frac{1}{N} \sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{u}_i)}. \tag{24}$$

В [16] представлено доказательство того факта, что InfoNCE loss максимизирует нижнюю границу взаимной информации.

Теорема 1. *$\mathcal{L}_{Pos}$ 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 в пределе при $N \to \infty$ лосс-функцию можно представить в виде:

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

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

$$\mathcal{L}_{\text{Pos}} = \mathbb{E}_{\substack{x\sim p\\ x_i\sim p_x^-}}\left[-\log\frac{\mathbb{E}_{x\sim p}f(x,c) - \tau^-\mathbb{E}_{x^-\sim p_x^-}f(x_i^-,c)}{\mathbb{E}_{x\sim p}f(x,c) + \big(N\tau^+ - \tau^-\big)\mathbb{E}_{x^-\sim p_x^-}f(x_i^-,c)}\right] =$$

$$= \mathbb{E}_{\substack{x\sim p\\ x_i^-\sim p_x^-}}\left[\log\frac{\mathbb{E}_{x\sim p}\frac{p(x|c)}{p(x)} + \big(N\tau^+ - \tau^-\big)\mathbb{E}_{x^-\sim p_x^-}\frac{p(x_i^-|c)}{p(x)}}{\mathbb{E}_{x\sim p}\frac{p(x|c)}{p(x)} - \tau^-\mathbb{E}_{x^-\sim p_x^-}\frac{p(x_i^-|c)}{p(x)}}\right] \approx$$

$$\approx \mathbb{E}_{\substack{x\sim p\\ x_i^-\sim p_x^-}}\left[\log\frac{\frac{1}{N+2}\left(\frac{p(x|c)}{p(x)} + \sum_{i=1}^{N}\frac{p(x_i^-|c)}{p(x_i^-)} + \frac{p(c|c)}{p(c)}\right) + \big(N\tau^+ - \tau^-\big)\frac{1}{N}\sum_{i=1}^{N}\frac{p(x_i^-|c)}{p(x_i^-)}}{\frac{1}{N+2}\left(\frac{p(x|c)}{p(x)} + \sum_{i=1}^{N}\frac{p(x_i^-|c)}{p(x_i^-)} + \frac{p(c|c)}{p(c)}\right) - \tau^-\frac{1}{N}\sum_{i=1}^{N}\frac{p(x_i^-|c)}{p(x_i^-)}}\right] \approx$$

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

$$= \mathbb{E}_{x\sim p}\left[\log\frac{\frac{p(x|c)}{p(x)} + N + 1 + N^2\tau^+ + 2N\tau^+ - N\tau^- - 2\tau^-}{\frac{p(x|c)}{p(x)} + N + 1 - (N+2)\tau^-}\right] =$$

$$= \mathbb{E}_{x\sim p}\log\left[1+\frac{N(N+2)\tau^+}{\frac{p(x|c)}{p(x)} + N\tau^+ + 1 - 2\tau^-}\right] = \mathbb{E}_{x\sim p}\log\left[1+\frac{N(N+2)\tau^+\frac{p(x)}{p(x|c)}}{1 + (N\tau^+ + 1 - 2\tau^-)\frac{p(x)}{p(x|c)}}\right] \geqslant$$

$$\geqslant \mathbb{E}_{x\sim p}\log\left[\frac{N(N+2)\tau^+\frac{p(x)}{p(x|c)}}{1 + (N\tau^+ + 1 - 2\tau^-)\frac{p(x)}{p(x|c)}}\right] =$$

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

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

При $N \to \infty$ получим:

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

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

Следовательно,

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

При условии того, что $\frac{p(x|c)}{p(x)}$ задаётся распределением данных в выборке и не зависит от N, получаем, что первое слагаемое в выражении справа при $N \to \infty$ больше нуля.

■

4 Вычислительный эксперимент

4.1 Классификация

В первом вычислительном эксперименте проводится сравнение $\mathcal{L}_{N-pair}$ (2) и $\mathcal{L}^N_{\text{Pos}}(f)$ (22) в задаче классификации изображений. В качестве датасета используется CIFAR10 [10], состоящий из 60000 цветных изображений размером 32x32 для классификации. Бейзлайн-модель: SimCLR [5] с энкодером ResNet18 [7] и N-pair loss (2), оптимизатор – Adam [9], learning rate 0.001, размер батча 256. Все модели тренируются 50 эпох и оцениваются линейным классификатором после получения эмбеддинга.

Сравнение результатов работы можно увидеть на рис.1-2. В качестве метрики берётся топ-1 и топ-5 accuracy:

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

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

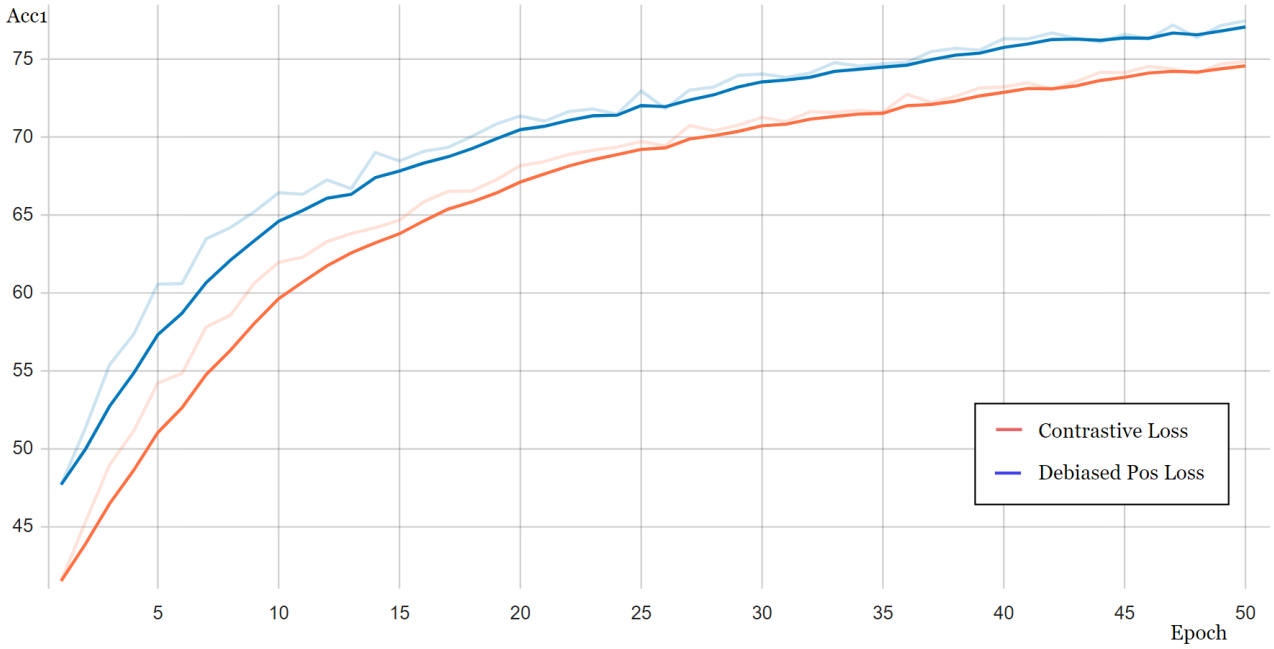

Рис. 1: Метрика acc1 классификации с использованием N-pair loss и DebiasedPos loss.

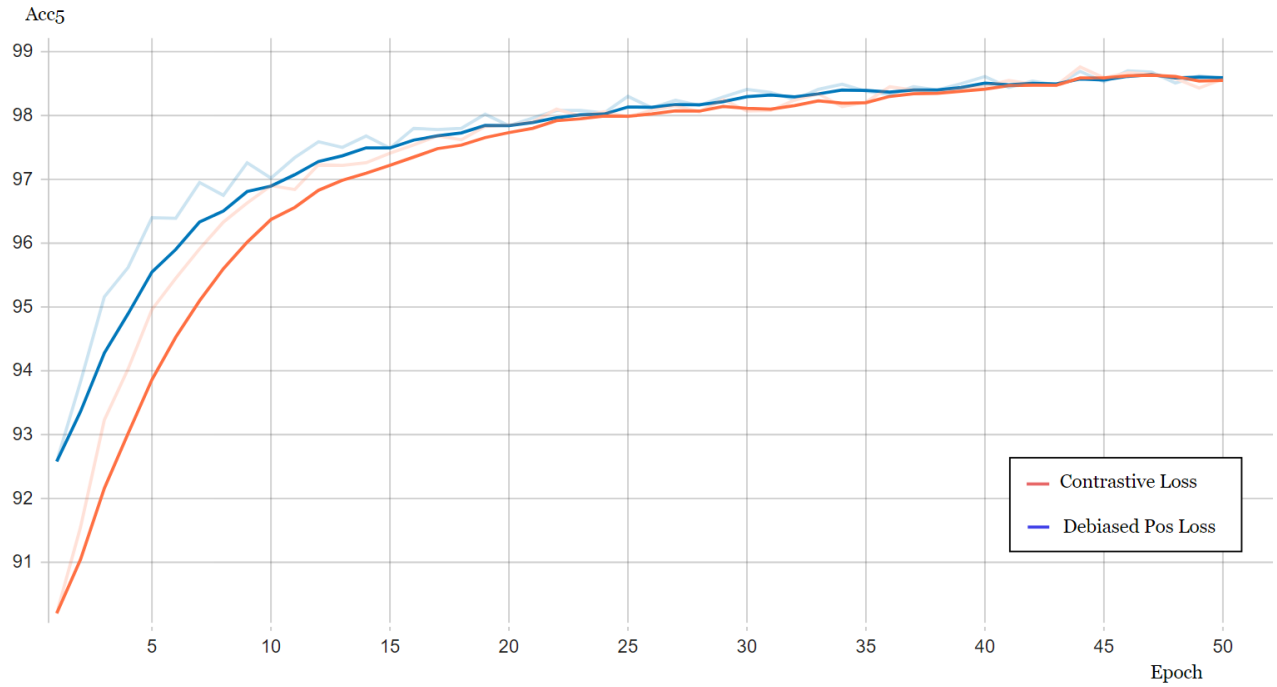

Рис. 2: Метрика acc5 классификации с использованием N-pair loss и DebiasedPos loss.

Как видно, модель, использующая $\mathcal{L}^N_{\text{Pos}}(f)$ имеет метрику топ-1 accuracy на 3.5% лучше. Топ-5 accuracy по истечении 50 эпох у моделей одинаковая, однако для модели с $\mathcal{L}^N_{\text{Pos}}(f)$ метрика растёт быстрее.

4.2 Искусственный эксперимент

Одна из задач данной работы – анализ пространства представления, порождённого предложенной лосс-функцией. Для этого генерируется стандартное нормальное распределение $\mathbf{z}$ в качестве изначальных данных, затем создаётся две модели нормализующего потока, которые каждый вектор $\mathbf{z}$ переводят в вектор, заданный некоторым другим распределением. В качестве таких распределений взяты moons и blobs из библиотеки sklearn. Полученные из одного и того же $\mathbf{z}$ вектора $\mathbf{a}$ и $\mathbf{b}$ подаются на вход энкодеру, состоящему из двух полносвязных слоёв.

Лосс-функция делится на две части: первая отвечает за сближение векторов, порождённых из одного и того же изначального вектора – это $\mathcal{L}_{N-pair}$, $\mathcal{L}_{Neg}$ или $\mathcal{L}_{Pos}$. Так как они используют косинусное расстояние, выходы энкодеров нормируются перед подсчётом лосс-функции. Вторая часть – приближение итогового распределения к нормальному. Так как

contrastive loss имеет несколько локальных минимумов и не обязательно каждый из них описывает нормальное распределение, для выхода каждого энкодера подсчитывается следующая функция потерь: из среднего по батчу и его дисперсии генерируется нормальное распределение, которое с помощью дивергенции Кульбака-Лейблера сравнивается со стандартным нормальным распределением.

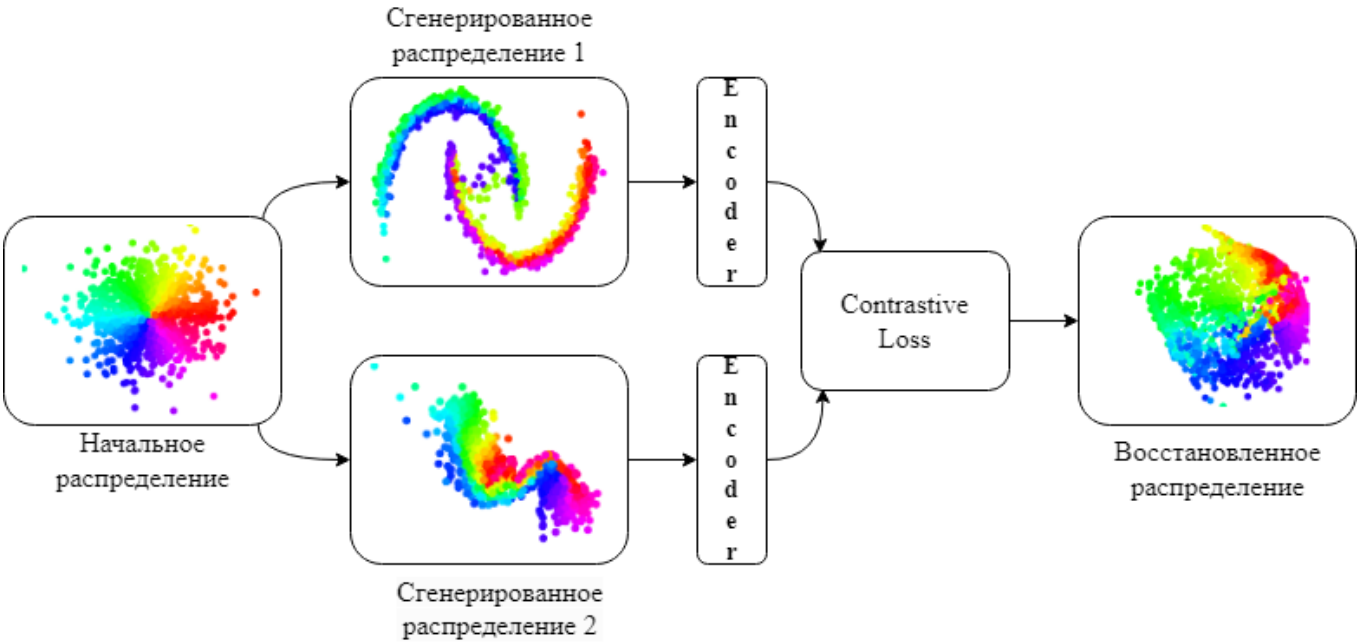

Рис. 3: Схема модели искусственного эксперимента. Точки одинакового цвета получены из одной и той же точки начального распределения.

4.3 VQA

В качестве большого эксперимента на реальных данных взята задача Visual Question Answering [1]. Датасет состоит из 204 721 картинок из MS COCO [13], 760 000 вопросов и около 10 миллионов ответов. Бейзлайн модель – TCL [18] с лосс-функцией $\mathcal{L} = \mathcal{L}_{cma} + \mathcal{L}_{imc} + \mathcal{L}_{lmi} + \mathcal{L}_{itm} + \mathcal{L}_{mlm}$. Все сравниваемые модели дообучаются с предложенной авторами [18] предобученной модели при замороженном визуальном энкодере 5 эпох на 100 000 вопросах и оценивается на 50 000 вопросах. В качестве метрики используется accuracy, которая равна 1, если ответ модели находится в предложенном авторами VQA списке из 10 ответов для каждого вопроса.

Новые модели создаются посредством замены $\mathcal{L}_{N-pair}$ в $\mathcal{L}_{cma}$ (5), $\mathcal{L}_{imc}$ (6), $\mathcal{L}_{lmi}$ (7) на $\mathcal{L}_{Pos}$ и $\mathcal{L}_{Neg}$. В качестве метрики берётся число попаданий

Таблица 1: Результаты VQA для $\mathcal{L}_{N-pair}$ и $\mathcal{L}_{Pos}$

	$\mathcal{L}_{N-pair}$	$\mathcal{L}_{Pos}$
Accuracy	0.67	0.69

ответа модели в список из 10 ответов, предоставленных составителями датасета:

$$Acc = \frac{1}{n}\sum_{i=1}^{n}[y_i \in y_i^{10}]$$

Результаты представлены в таблице 1.

Рис. 4: Пример работы модели в VQA-задаче. Вопрос: «What is the child eating?» Ответ модели с $\mathcal{L}_{Pos}$: «donut».

5 Заключение

В данной работы был проведён анализ смещения положительного и отрицательного распределения в задаче контрастного обучения на примере классификации изображений, искусственного двумерного эксперимента и VQA. Была предложена лосс-функция, учитывающая шум при сэмплировании положительных элементов, доказана её сходимость к $\mathcal{L}_{N-pair}$ и свойство максимизации правдоподобия между сравниваемыми положительными элементами.

Список литературы

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] 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.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[5] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning, 2020.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[8] 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.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[10] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[11] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

[12] 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.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[14] 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.

[15] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Neural Information Processing Systems*, 2016.

[16] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[17] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018.

[18] 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.