Аннотация

Контрастное обучение – распространённый подход в сфере обучения без учителя. Данный метод заключается в попарном сравнении представлений объектов выборки для восстановления распределения в пространстве представлений. При этом похожие представления находятся близко, а отличающиеся – далеко. Однако исходное распределение и способ порождения данных неизвестны, а функции потерь имеют несколько локальных минимумов, не все из которых соответствуют истинному распределению. В данной работе в качестве решения данных проблем приводятся функции потерь, устраняющие смещение распределений вследствие наличия ложноотрицательных и ложноположительных пар в разметке. Предлагается несмещённая модель контрасного обучения, исследуются её свойства. Качество полученного представления оценивается в задаче классификации и в задаче VQA, для которых предложенная модель показала результат лучше, чем модель, не учитывающая смещение. Также проведится эксперимент на искусственной выборке из двумерного пространства, в котором проверяется качество восстановления исходного распределения предложенной моделью.

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

Содержание

1 Введение

Методы обучения без учителя не требуют предварительной разметки данных, а, значит, для них легче найти большие по объёму выборки для предобучения универсальную модель, которая далее на небольшом датасете подстраивается под требуемую задачу.

Самообучение – класс методов, позволяющих модели обучаться на датасетах без явной разметки. Вместо этого используются некоторые наблюдаемые характеристики выборки.

В частности, *обучение представлениям* – это процесс отображения необработанных входных данных в пространство векоторов-признаков или тензоров с целью найти и извлечь полезные закономерности, которые будут полезны для предсказании целевых значений.

Первые модели машинного обучения использовались для принятия решений на основе извлеченных с помощью предобработки признаков. Увеличение объема доступных вычислений и размеченных датасетов позволило перейти от использования средств извлечения признаков, разработанных вручную, к моделям, автоматически выделяющим особенности выборки. В результате исследования также сместились с разработки признаков на разработку архитектуры модели, которая самостоятельно может создавать представление данных.

Задача поиска хорошего представления достаточно сложная. Согласно [12], существует несколько принципов, на которых строится обучение представления: выразительность и способность кодировать экспоненциальное число конфигураций в отличие от методов наподобие one-hot, абстрактность и отсутствие сильной зависимости от локальных возмущений, непротиворечивость и обеспечение наглядности представления. Одно из решений – *контрастное обучение* – 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 распределение входной выборки, при констрастном обучении представление изучается путем сравнения пар элементов. Пусть вектор $\mathbf{x}$ некоторого объекта принят за основной. Тогда вектор схожего объекта $\mathbf{x}^+$ – положительный элемент. Он должен располагаться как можно ближе к $\mathbf{x}$ в пространстве эмбеддингов ввиду гипотезы о генерации данных из распределения, в котором расстояние между схожими элементами минимально. А вектор отличного объекта $\mathbf{x}^-$ – как можно дальше, как негативный элемент. Основной сложностью использования такого подхода является правильный подбор негативных и позитивных элементов при отсутствии разметки в данных.

Одни из самых распространённых мер различия элементов: *Triplet loss* [15] и его модификация для N негативных примеров – *Multi-Class N-pair loss* [16], а также *Information Noise-Contrastive Estimation* или *InfoNCE* [17] 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. Обучение модели направлено на уменьшение расстояния между эмбеддингами $\mathbf{x}$ и $\mathbf{x}^+$ и увеличение расстояния между эмбеддингами $\mathbf{x}$ и $\mathbf{x}^-$:

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

В задачах распознавания изображений распространённым методом выбора положительной пары является аугментация основного изображения, а в качестве отрицательных пар берутся все остальные $N - 1$ элементов

в батче. Для случая $N-1$ негативных векторов существует обобщение $\mathcal{L}_{Triplet}(\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\})$:

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

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 обучения сравнениями имеют хороший результат в формировании визуальных представлений. Например, в [4] представлен фреймворк SimCLR, значительно превосходящий предыдущие методы самообучения. 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 вектора из сформированного пространства представлений в функцию потерь входит его проекция на некоторое другое пространство.

В методах самообучения отсутствие истинных меток приводит к необходимости случайного выбора отрицательных экземпляров $\mathbf{x}^-$. Однако такой подход приводит к смещению, если $\mathbf{x}^-$ на самом деле достаточно похож на $\mathbf{x}$, например, фотографии одного и того же объекта с разных ракурсов. В таком случае смещение приводит к отдалению близких элементов

в пространстве эмбеддингов и, как следствие, к значительному снижению качества [9].

В [5] предложен несмещённая контрастная функция потерь. Учитывается N отрицательных и M положительных элементов из выборки. $L_{\text{Neg}}^{N,M}(f)$ – функция потерь, корректирующая смещение ввиду наличия ложноотрицательных экземпляров. 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 в пространстве эмбеддингов.

2 Предыдущие исследования

Основная задача данного исследования – создание функции потерь, учитывающей смещение в данных и анализ её способности восстанавливать исходное распределение, из которого были порождены данные, в пространстве эмбеддинов. 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 подобным образом. Одна из рассматриваемых задач – классификация на неразмеченных данных. В этом случае большое влияние на качество результата оказывает подбор для $\mathbf{x}$ отрицательнх пар. Ввиду того, что в качестве положительных пар берутся различные аугментации данного изображения, а в качестве отрицательных – все остальные картинки в батче, появляется вероятность получить ложноотрицательные пары, так как есть вероятность, что не все изображения в выборке имеют класс, отличный от класса элемента $\mathbf{x}$. В [5] в функции потерь вероятность получить отрицательную пару выражается через полную вероятность и вероятность получить положительную пару. Тем самым приближается неизвестное распределение отрицательных элементов в выборке.

В [13] и [19] рассматривается применение обучения сравнениями в задаче сопоставления картинок и текстов для предобучения. В [13] 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 [19], в отличие от [13], также учитывается проблема того, что обычно в изображении хранится больше информации, чем описано в тексте. В [19] применяется локальное выравнивание для максимизации взаимной информации между локальным и глобальным представлениями. Итоговая функция потерь в данном фреймворке состо-

ит 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 энкодеров на пространство, в котором применяется лосс-функция. Для выбора отрицательных текстовых элементов используется идея, берущая начало с *банка памяти* из [18], который для уменьшения числа вычислений эмбеддингов хранит представления с предыдущих итераций, что увеличивает затраты по помяти, но значительно уменьшает время работы модели. Эта идея имеет развитие в [7], где используются эмбеддинги не самой модели, а её версии с моментумом для избавления от шума, являющегося большой проблемой для больших датасетов, собранных без разметки из интернета. 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:

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

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

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. Такое выравнивание используется, так как объекты разной модальности зачастую не могут полностью описать друг друга. В таком случае необходимо 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, но и локальную, в итоговую функцию потерь в [19] 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 текстов, используется задача бинарной классификации предсказания, положительна пара или отрицательна, где в качестве таргета используется результат применения функции потерь из кросс-модального выравнивания, а в качестве предсказания – выход полносвязного слоя мультимодального 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, распространённого для

больших датасетов, и улучшения гибкости модели в использовании синонимов. 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 [19]:

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

Обоснование применения InfoNCE заключается в том, что он связан с нижней границей взаимной информации, которую тяжело посчитать напрямую. Уменьшение InfoNCE увеличивает взаимную информацию внутри положительной пары. В [17] представлено доказательство, использующее связь лосса с MINE [3].

3 Постановка задачи

3.1 Несмещённая функция потерь

Пусть $\mathcal{X}$ – предметное пространство. Обучение сравнениями 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» N-pair лосс при условии наличия N негативных элементов выглядит 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 достаточно 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 функции потерь в модели SimCLR позволило смоделировать смещение при наличии в выборке ложноотрицательных элементов и увеличить точность результата. В данной работе будет произведено её сравнение с функцией потерь, в которой вместо $p(\mathbf{x}^-)$ будет выражен $p(\mathbf{x}^+)$.

3.2 Ложноположительные элементы

Чрезмерная аугментация делает положительную пару достаточно далёкой друг от друга и приводит к появлению ложноположительных объектов и смещению оценки. При недостатке вычислительных ресурсов зачастую при аугментации из изображения вырезается некоторая часть, которая может сильно отличаться от второй аугментации. Чем меньше размер вырезанного изображения, тем больше вероятность появления ложноположительной пары. Распределение положительных элементов в рамках функции потерь представляется таким образом:

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

Лемма 1. *При $N \to \infty$ несмещённая функция потерь стремится к функции потерь, учитывающей наличие ложноположительных элементов:*

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

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. Для того, чтобы внести предел под знак математического ожидания, можно применить теорему Леви о монотонной сходимости. Для неё требуется, чтобы последовательность положительных измеримых функций под знаком математического ожидания была монотонно возрастающей. Тогда пусть $h_N(\mathbf{x}) = -\log \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^N e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}}$ – измеримая, как композиция измеримых функций.

Так как ввиду положительности экспоненты

$$\frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \leqslant 1,$$

то последовательность $\{h_N(\mathbf{x})\}_{N=1}^{\infty}$ состоит из положительных функций.

$$\frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)}} \geqslant \frac{e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \sum_{i=1}^{N} e^{f(\mathbf{x})^T f(\mathbf{x}_i^-)} + e^{f(\mathbf{x})^T f(\mathbf{x}_{N+1}^-)}}$$

Следовательно, так как логарифм – монотонная функция, а в математическом ожидании он умножается на -1, последовательность $\{h_N(\mathbf{x})\}_{N=1}^{\infty}$ монотонно возрастает:

$$0 \leqslant h_N(\mathbf{x}) \leqslant h_{N+1}(\mathbf{x})\}\}$$

Так как $h = -\log \frac{\mathbb{E}_{\mathbf{x}^+ \sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)}}{\mathbb{E}_{\mathbf{x}^+ \sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + N \mathbb{E}_{\mathbf{x}^- \sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}}$ – измеримая функция, то, учитывая доказательство выше, можно утверждать, что условия теоремы Леви выполнены. Тогда:

$$\lim_{N \to \infty} \mathbb{E} h_N(\mathbf{x}) = \mathbb{E} h(\mathbf{x})$$

Переходя к исходным обозначениям:

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

Используя 17 и линейность матожидания, получим:

$$\mathbb{E}_{\mathbf{x}'\sim p} e^{f(\mathbf{x})^T f(\mathbf{x}')} = \tau^+ \mathbb{E}_{\mathbf{x}^+\sim p_x^+} e^{f(\mathbf{x})^T f(\mathbf{x}^+)} + \tau^- \mathbb{E}_{\mathbf{x}^-\sim p_x^-} e^{f(\mathbf{x})^T f(\mathbf{x}^-)}$$

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

■

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

Введём обозначение предела в лемме $\tilde{\mathcal{L}}^N_{\text{Pos}}(f)$:

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

Полученная функция потерь при конечном N:

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

где эмпирические оценки матожиданий:

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

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

В [17] представлено доказательство того факта, что InfoNCE loss максимизирует нижнюю границу взаимной информации между положительной парой, где c – контекст, x – элемент выборки:

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

Для того, чтобы показать, что при уменьшении $\mathcal{L}_{\text{Pos}}^{N}$ также увеличивается нижняя граница взаимной информации, необходимо доказать следующую теорему.

Теорема 1. *$\mathcal{L}_{Pos}$ максимизирует нижнюю границу взаимной информации, то есть:*

$$I(x, c) \geqslant A - \mathcal{L}_{Pos}$$

где A положительно.

Доказательство. Так как рассматриваемая функция потерь – это категориальная кросс-энтропия классификации положительного сэмпла корректно, то оптимальная для него вероятность того, что x_i – положительный элемент:

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

Функцию потерь, порождённую через кросс-энтропию, 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:

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

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

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

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

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

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

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

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

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

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, первое слагаемое в выражении справа при $N \to \infty$ больше нуля.

■

4 Вычислительный эксперимент

В работе проводится 3 вычислительных эксперимента. Первый – эксперимент на искусственных данных в двумерном пространстве для визуализации и анализа возможности предложенного метода восстанавливать начальное распредение в пространстве эмбеддингов. Второй эксперимент приводится для сравнения трёх рассмотренных в работе функций потерь на классической задаче классификации изображений на распространенном датасете. Третий эксперимент – сравнение предложенного метода и $\mathcal{L}_{\text{N-pair}}$ для предобучения мультимодальной модели, соединеняющей тексты и изображения в общем скрытом пространстве для последующего использования в задаче ответа на вопросы по изображению.

4.1 Эксперимент на искусственных данных

Одна из задач данной работы – анализ пространства представления, порождённого предложенной функцией потерь. Для этого проводится экперимент на искуссвенных данных. Двумерное распределение в изначальном пространстве $\mathbf{z}$ генерируется из стандартного нормального распределения $\mathcal{N}(\mathbf{0}, I)$. Затем две модели нормализующего потока Real NVP [6] каждый вектор $\mathbf{z}$ переводят в вектор, заданный некоторым другим распределением. Таким образом генерируются две модальности данных, порождённые из одного начального распределения аналогично предположению о том, что изображения и подписи к ним порождены из близких точек в начальном пространстве. В качестве новых распределений взяты луны и 5 кругов из библиотеки sklearn. Полученные из одного и того же $\mathbf{z}$ вектора $\mathbf{a}$ и $\mathbf{b}$ подаются на вход двум различным энкодерам, состоящим из двух полносвязных слоёв.

Функция потерь 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 ненормированного 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.

Модель обучается 50 эпох, в батче 32 вектора, оптимизатор – Adam [10], шаг обучения 0.001.

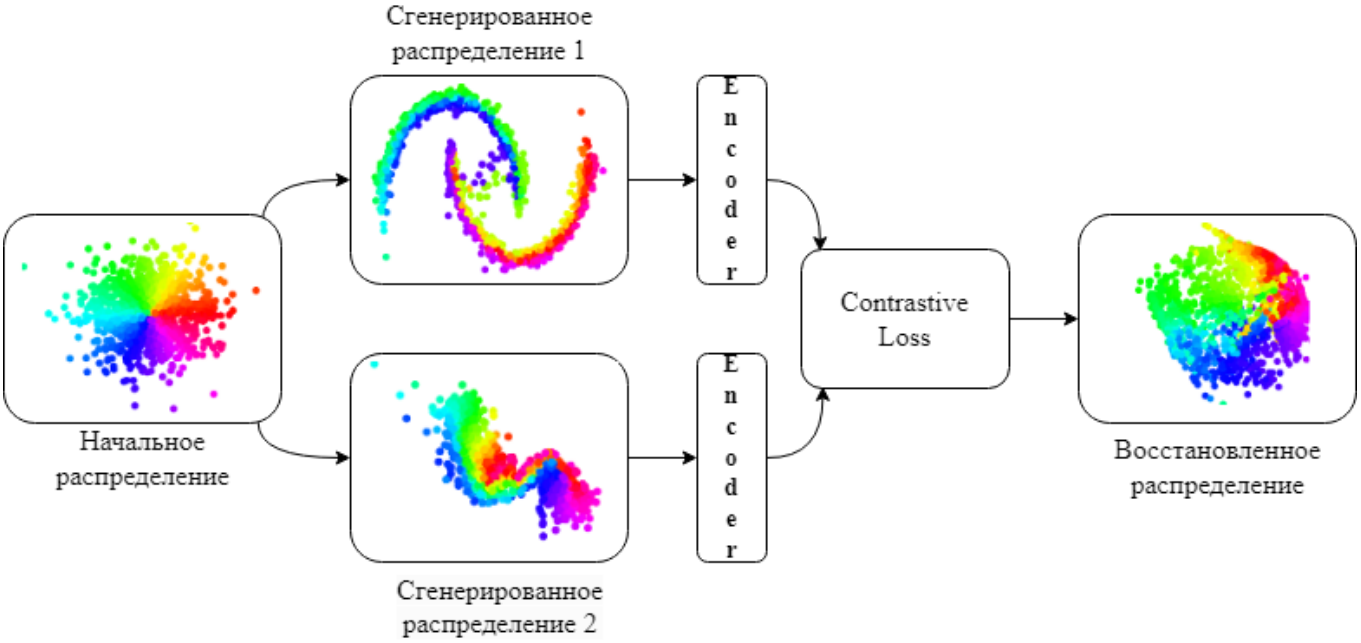

Рис. 1: Схема модели искусственного эксперимента. Точки одинакового цвета получены из одной и той же точки начального распределения.

На рис.1 представлена визуализация работы модели. Локальный минимум представленной функции потерь достаточно хорошо восстанавливает форму начального распределения. Также можно заметить артефакты, характерные для сгенерированных распределений. Например, более плотное скопление для точек красного и розового цвета и менее плотное – для точек зелёного цвета. На практике не получится полностью смоделировать начальное распределение ввиду шумов и характерных особенностей для каждой

модальности, однако приблизить к нему восстановленное распределение можно.

В качестве метрик качества приближения векторов в пространстве представления используется MAE и косинусное расстояние:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \|x_i - y_i\|$$

$$Cosine = \frac{1}{n}\sum_{i=1}^{n} \frac{\langle x_i, y_i \rangle}{\|x_i\| \cdot \|y_i\|}$$

Таблица 1: Метрики качества приближения векторов в пространстве представления для $\mathcal{L}_{N-pair}$ и $\mathcal{L}_{Pos}$

	До модели	$\mathcal{L}_{N-pair}$	$\mathcal{L}_{Pos}$
MAE	1.508	0.616	**0.564**
Cosine	0.080	0.897	**0.911**

4.2 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 [11], состоящий из 60000 цветных изображений размером 32x32 для классификации. Базовая модель: SimCLR [5] с энкодером ResNet18 [8] и N-pair loss (2), оптимизатор – Adam [10], шаг обучения 0.001, размер батча 256. Все модели тренируются 50 эпох и оцениваются линейным классификатором после получения эмбеддинга.

Сравнение результатов работы можно увидеть на рис.3 и в таблице 2. В качестве метрики берётся топ-1 и топ-5 accuracy:

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

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

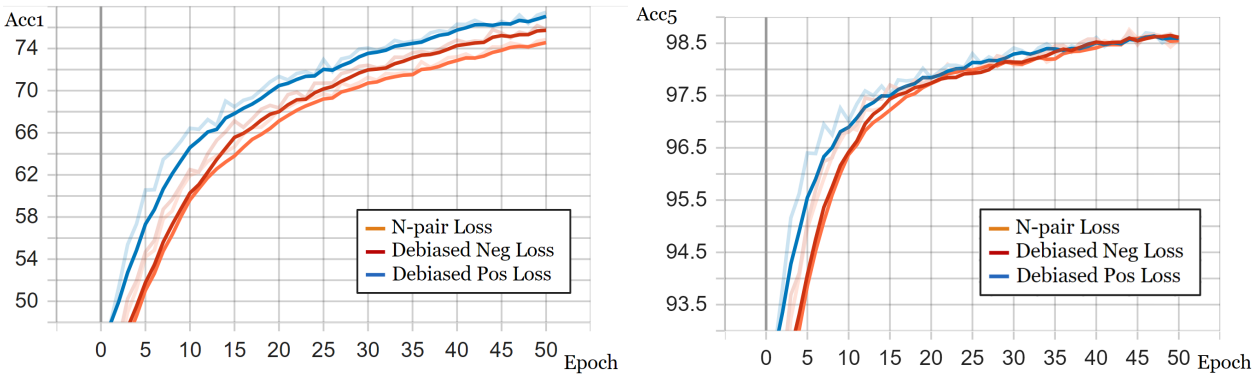

Рис. 2: Метрики acc1 и acc5 классификации с использованием N-pair loss, DebiasedNeg loss и DebiasedPos loss.

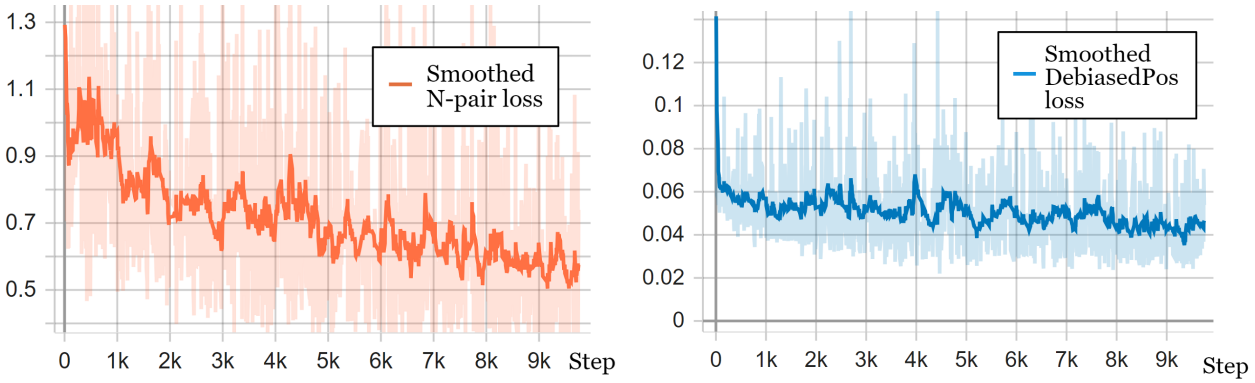

Рис. 3: Графики функции потерь для классификации с использованием N-pair loss и DebiasedPos loss.

Таблица 2: Результаты классификации для $\mathcal{L}_{N-pair}$, $\mathcal{L}_{Neg}$ и $\mathcal{L}_{Pos}$

	$\mathcal{L}_{N-pair}$	$\mathcal{L}_{Neg}$	$\mathcal{L}_{Pos}$
Acc1	74.84	75.81	**77.45**
Acc5	98.56	98.56	**98.58**

Модель, использующая $\mathcal{L}^N_{\text{Pos}}(f)$ имеет метрику топ-1 accuracy на 2.6% лучше. Топ-5 accuracy по истечении 50 эпох у моделей одинаковая, однако для модели с $\mathcal{L}^N_{\text{Pos}}(f)$ метрика растёт быстрее.

4.3 VQA

В качестве большого эксперимента на реальных данных взята задача ответа на вопросы по изображению [1]. Датасет состоит из 204 721 картинок из MS COCO [14], 760 000 вопросов и около 10 миллионов ответов. Бейзлайн модель – TCL [19] с функцией потерь $\mathcal{L} = \mathcal{L}_{cma} + \mathcal{L}_{imc} + \mathcal{L}_{lmi} + \mathcal{L}_{itm} + \mathcal{L}_{mlm}$. Все сравниваемые модели дообучаются с предложенной авторами [19] 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 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 3.

Таблица 3: Результаты VQA для $\mathcal{L}_{N-pair}$ и $\mathcal{L}_{Pos}$

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

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] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2016.

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

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

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

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

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

[19] 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.