# Нейросетевой подход к обучению векторных представлений мультимодальных данных на основе угловых аддитивных функций потерь

Черникова Полина Георгиевна
ВМК МГУ
p.chernikova@yandex.ru

Воронцов Константин Вячеславович
ВМК МГУ
vokov@forecsys.ru

2023

Аннотация

В данной работе рассматривается задача обучения метрического представления мультимодальных данных. Для ее решения предлагается новый подход обучения с учителем на основе сферического классификатора и аддитивных угловых функций потерь. Предлагаемый подход состоит из двух основных компонентов: 1) двухуровневой архитектуры на базе трансформеров; 2) аддитивной угловой функции потерь, напрямую оптимизирующей сферическое расстояние между объектами. Эксперименты показали, что данный подход позволяет ускорить обучение, получить более разделимые и геометрически интерпретируемые векторные представления по сравнению с классическими подходами. Предложенная архитектура легковесна, эффективно скалируется при увеличении количества данных и модальностей, не требуя перебора комбинаторного числа пар, как контрастивное обучение.

Ключевые слова: Векторное представление · Мультимодальные данные · Аддитивная угловая функция потерь · Нейронные сети

## 1 Введение

Задача обучения векторных представлений - это задача машинного обучения, в ходе которого алгоритмы извлекают значимые закономерности из исходных данных и создают представления, которые легче понять и обработать. Выученные пердставления используются для широкого класса задач, таких как: поиск, рекомендации, кластеризация, классификация и т.д. Zhang et al. [2020a].

Важной частью данной задачи является обучение на мультимодальных данных, то есть данных из разных источников, разной природы Ngiam et al. [2011]. В реальном мире большинство задач основаны на мультимодальных данных: рекомендации в e-commerce Bai et al. [2023], медицине Huang et al. [2021], робототехника Islam and Iqbal [2021] и т.д. Кроме того, добавление мультимодальных данных может значительно учлушить результаты решаемой задачи Zhang et al. [2019].

При построении векторных представлений мультимодальных данных основной трудностью и темой для исследований является задача смешения (фьюжена) данных - интеграция информации, извлеченной из каждой модальности в единое компактное представление. Методы фьюжена можно разделить в зависимости от стадии, на которой происходит слияние данных, на ранние и поздние. При раннем фьюжене данные соединяют до применения к ним модели, при позднем - данные каждой модальности подаются на вход соответсвующим моделям, и уже их выходные признаковые представления конкатенируются. Последние исследования [Zhang et al., 2020b] используют промежуточный фьюжен, когда слияние происходит на нескольких слоях глубокой сети. Промежуточный фьюжен может быть реализован

- простыми операциями конкатенирования, взвешенных сумм (полносвязный слой) или их суперпозиции Nikzad-Khasmakhi et al. [2021];
- с помощью механизмов внимания, которые зачастую представляют из себя взвешенную суму набора векторов с весами, которые генерируются динамически на каждом шаге маленькими моделями;
- с помощью билинейного пулинга, который создает совместное представление для векторов двух модальностей, вычисляя их внешнее произведение

Существует 3 основных нейросетевых фреймфорка для построения мультимодальных векторных представлений [Guo et al., 2019]:

- совместное представление (joint representation), выучивающее общее семантическое пространство для всех модальностей
- координированное представление (coordinated representation), выучивающее отдельные представления для каждой модальности, которые координированы за счет ограничений (констрейнов)
- система энкодер-декодер, которая переводит одну модальность в другую, сохраняя их семантику согласованной

Большинство современных нейросетевых методов обучения векторных представлений мультимодальных данных для доменных задач поиска и рекомендаций сегодня реализованы на основе фреймворка joint representation и используют контрастивное обучение Bai et al. [2023], Dong et al. [2022], Huang et al. [2019]. Этот подход показывает результаты, сравнимые с SOTA, однако имеет ряд недостатков. Так для контрастивного обучения требуется подготовка большого количества положительных и отрицательных пар объектов, которое возрастает комбинаторно с ростом размера датасета, что приводит к долгому и ресурсозатратному обучению. В основном в качестве модальностей используется пара: изображение - текст, при этом игнорируется важность дополнительной информации из табличных данных, категориальных признаков и взаимосвязей. Однако, как показано во множестве работ .

В данной работе предлагается новый подход обучения метрических векторных представлений мультимодальных данных, использующий совместное представление модальностей на сфере и угловех аддитивные функций потерь. Похожий подход на основе сферического классификатора и угловых аддитивных функций потерь зарекомендовал себя в задаче распознания лиц Deng et al. [2019]. Он не только ускорил обучение на больших датасетах, но и повысил дискриминативную способность векторных представлений признаков в сравнении с контрастивным обучением. Предложенный фреймворк состоит из двух основных компонентов: двухуровневой архитектуры на базе трансформеров и аддитивной угловой функции потерь, напрямую оптимизирующей сферическое расстояние между объектами. Этот фреймворк позволяет эффективно увеличивать количество данных и модальностей, не требует перебора комбинаторного числа пар, как контрастивное обучение, и обучает компактные, разделимые векторные представленяи, которые хорошо подходят для задач поиска, кластеризации и рекомендаций.

## 2 Задача обучения метрических векторных представлений на мультимодальных данных

Цель обучения метрических векторных представлений мультимодальных данных состоит в том, чтобы выучить общее пространство представлений, которое отражает признаки каждой модальности и внутрениие взаимосвязи между ними. Полученные в этом пространстве представления разделимы и репрезентативны. Представления похожих объектов, то есть объектов относящихся к одной гипотетической категории должны быть близки. Формализуем данную постановку. Пусть у нас есть набор данных $D$ с модальностями , где $|M| = n$.

Пусть $d_i \in D$ - набор данных из $n$ модальностей, описывающий $i$-й объект, где $i \in \hat{1,n}$; $Y$ - множество множество гипотетических категориий. В общем случае это множество не ограничено. Каждой категории может принадлежать сразу несколько наборов данных $d_i$. Надо найти такое отображение $f : D \to R$, что:

$$\underset{i=j}{\text{sim}}(r_{y_i}, r_{y_j}) < \underset{i\neq j}{\text{sim}}(r_{y_i}, r_{y_j}) \forall i, j \in \mathbf{N}, r \in R$$

где:

- $r_{y_i}$ - вектор из R, принадлежащий категории $y_i$
- $\text{sim}(A, B) = \frac{A*B}{||A||*||B||}$ - косинусная близость векторов

Конкретно в нашей задаче будем использовать следующие n = 3 модальности:

1. $T$ - текст
2. $C$ - категориальные
3. $V$ - табличные данные

Так как у нас нет информации о количестве всех классов (категорий), будем использовать прокси метрику ROC-AUC на бинарной клссификации принадлежности пары к одному классу. В качестве уверенности модели будем использовать косинусную близость между объектами пары.

$$\text{ROC-AUC} = 1 - \int_0^1 \text{FRR}\, d\text{FAR}$$

, где

$\text{FRR} = \frac{\text{False Negatives}}{\text{False Negatives+True Positives}}$

$\text{FAR} = \frac{\text{False Positives}}{\text{False Positives+True Negatives}}$

## 3 Решение

### 3.1 Архитектура

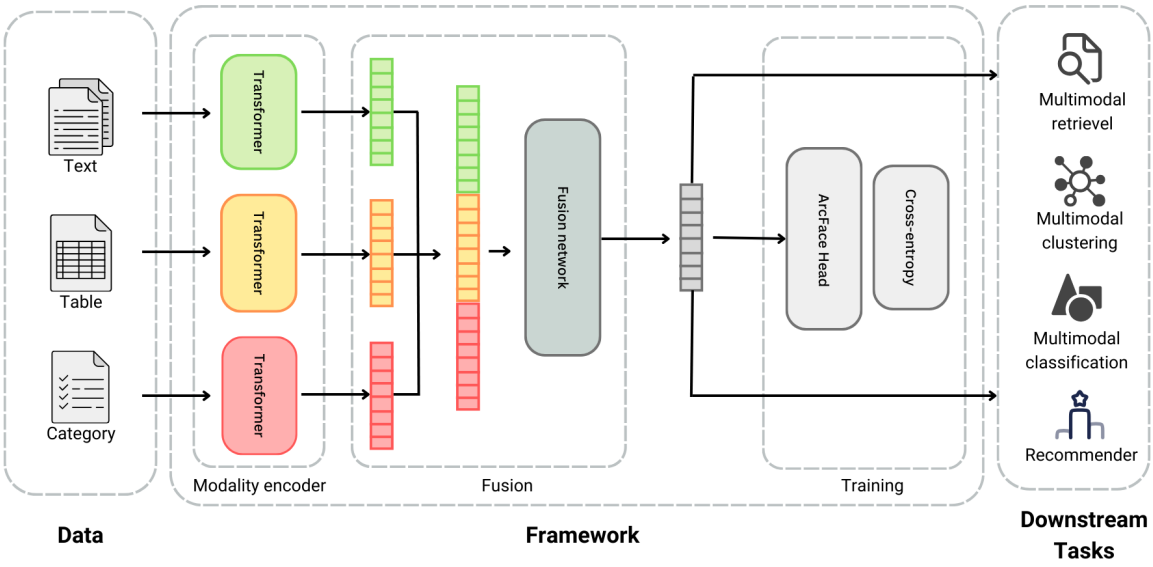



Рис. 1: Модель на вход принимает данные трех модальностей: текст, табличные данные, категориальные признаки. Каждая модальность представляется вектором признаков, полученным с помощью трансформера, специфичного для этой модальности. Полученные вектора конкатенируются и подются на вход смешивающей сети. Смешивающая сеть обучается с помощью сферической головы ArcFace и кросс-энтропии. На этапе инференса полученные в смешивающей сети компактный и дискриминативные представления можно использовать в задачах мултимодального поиска, кластеризации, классификации и построения рекомендаций.

В качестве решения предлагается фреймворк, архитектура которого представлена на рис. 1. Для каждой модальности данных используется свой модально-специфичный трансформер. Так для текста на английском языке мы использовали Sentence BERT Reimers and Gurevych [2019]. Для табличных и каткгориальных данных могут быть использованы предобученные трансформерные модели для структурированных и текстовых данных на основе BERT, например, TaBERT Yin et al. [2020] или Tabbie Iida et al. [2021]. Однако в данном случае мы использовали простейшее кодирование признаков: нормализацию вещественных признаков и one-hot encoding для категориальных. Полученные векторные представления конкатенируются и подаются на вход смешивающей нейронной сети (Fusion network), которая в базовом виде состоит из одного полносвязного слоя, проектирующего мультимодальный вектор в общее пространство признаков. Смешивающая сеть обучается с помощью головы сферического классификатора ArcFace Deng et al. [2019] и кросс-энтропии. Идея состоит в том, чтобы отобразить представление на гиперсферу и с помощью аддитивной угловой функции потерь (additive angular margin loss) раздвинуть векторы признаков, принадлежащие разным классам.

Формально сферическая классификационная голова работает следующим образом:

- Пусть на выходе полносвязного слоя смешивающей сети мы получили вектор $x \in \mathbf{R}^d$. Нормализуем его, чтобы отобразить на сферу: $||x||_2 = 1$
- Нормализованный вектор может быть представлен следующим образом:

$$\frac{W^T x}{||W^T x||_2}$$

  , где $W \in \mathbf{R}^{d \times n}$ - матрица весов полносвязного слоя, $n$ - число классов (число категорий в нашей постановке), $x$ - сконкатенированный модальности вектор, который мы подаем на вход полносвязному слою нашей смешивающей сети.
- Логиты сети для класса $y_i$ могут быть получены как скалярное произведение матрицы весов на вектор признаков, полученных на выходе полносвязного слоя: $W_{y_i}^T f = ||W_{y_i}||||f|| \cos\theta_i$, где $W_{y_i}$ - $i$-й столбец матрицы W, соответсвующий классу $y_i$, $\theta_i$ - угол между весом $W_{y_i}$ и признаков $f_i$.
- С помощью l-2 нормализации, аналогично $x$, приведем $||W_i|| = 1$, а $||W_i|| = 1$ скалируем до длины радиуса гиперсферы $s$, на которую отображаем признаки. Такая нормализация признаков и весов приводит к тому, что теперь предсказания зависят только от угла между признаками и весами.
- Таким образом выученные векторные представления признаков распределены на гиперсфере радиуса $s$.
- Чтобы раздвинуть признаки, принадлежащие разным классам, введем штрафную добавку $m$ между $x_i$ и $W_{y_i}$. Таким образом получим аддитивную угловую функцию потерь, штрафующую геодезическое расстояние между классами:

$$L = \log\left(\frac{e^{s(\theta_{y_i}+m)}}{e^{s(\theta_{y_i}+m)} + \sum_{j \neq y_i} e^{s\theta_j}}\right)$$

  где
  - $s$ радиус гиперсферы или скалирующий коэффициент,
  - $\theta_i$ - векторное представление $i$-го признака в углвой форме
  - $y_i$ - реальныя метка класса (категории) $i$-го вектора признаков.
  - $m$ - угловая добавка

В геометрической интерпретации сефрическая классификационная голова отображает вектора признаков на сферу таким образом, чтобы признаки одного класса концентрировались на сфере под одним углом. Аддитивная угловая функция потерь с помощью углового сдвига раздвигает признаки от границы решений (decision boundery), тем самым обучает разделимые векторные представления.

Предложенная архитектура легковесна.

### 3.2 Данные

Для проведения экспериментов и тестирования модели, предложенной в качестве решения, был взят набор Google Local Reviews картах 2021, содержащий информацию о компаниях и отзывах о них на

| № | Текст | Таб. числовые | Таб. категории |
|---|---|---|---|
| 1 | Название компании | Долгота | Категория компании (ниша) |
| 2 | Адрес | Широта | Уровень цен |
| 3 | Описание компании | Средний рейтинг | Варианты обслуживания |
| 4 | Часы открытия | | Серы безопасности |
| 5 | | | Доступность |
| 6 | | | Планировка |
| 7 | | | Способы оплаты |

Таблица 1: Структура признаков компании

Google картах. Этот датасет использовался в похожих работах про мультимодальность и контрастивное обучение Li et al. [2022], Yan et al. [2023].

В качестве датасета для теста была выбрана часть исходного датасета - метаданные о всех компаниях в Калифорнии. Эти данные содержали информацию о 515,961 компаниях. Набор данных о компании содержал признаки 3 модальностей (см. таб. 1).

#### 3.2.1 Обработка данных

Мы убрали из датасеты те компании, у которых не было текстового описания, чтобы не терять модальность полностью в условиях малого числа признаков. Название и адрес не использовались. Часы открытия были переведены в минуты и стали численным признаком. Все численные признаки были нормализированы. Категориальные признаки изначально пердставляли из себя список категорий, который мы перевели с помощью label encoding в множество новых признаков. Далее все категориальные признаки были закодированы one-hot encoding.Категория компании не использовалась в кодировании, так мы определили ее как категорию, метку класса. Текст описания компаний закодировали с помощью трансформерной модели sentence BERT. Таким образом было получена 3 вектора, соответсвующие текстовым, числовым и категориальным признакам. В итоге мы получили датасет из 105,365 компаний.

Мы разбили его на обучающую и тестовую в выборку в соотношении 70:30, тестовую выборку разбили еще на тест и валидацию в соотношении 90:10. Разбиение на обучающую и тестовую выборку было произведено таким образом, чтобы категории объектов в этих выборках не пересекались.

#### 3.2.2 Разметка пар

В формальной постановке задачи было оговорено наличие категории у каждого набора вектора. В нашем случае категория - это категория бизнеса. В нашем обучающем датасете получилось 768 уникальных категорий.

Для замера нашей основной метрики ROC-AUC необходимо разбиение на положительные и отрицательные пары, где положительной парой объектов будем считать объекты, относящиеся к одной категории. Тестовый датасет был разбит на пары таким образом, чтобы на каждую категорию было 20 положительных и 20 отрицательных пар.

## 4 Эксперименты

### 4.1 Бейзлайн

В качестве бейзлайна мы взяли датасет, который состоял только из текстовой модальности, закодированной sentence BERT. ROC-AUC только на текстовой модальности равен 0.93 рис. ??. График распределения межклассовой и внутриклассовой косинусной близости векторов на рис. 2a демонстрирует большой разброс скоров близости объектов из разных классов, вплоть до 0.85.

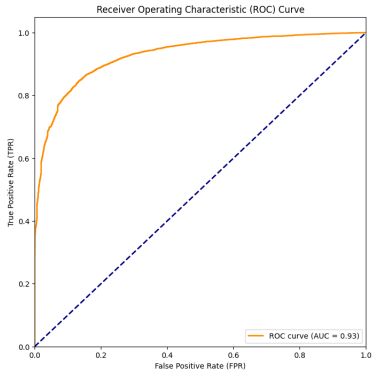
### 4.2 Добавление модальностей: конкатенация

Добавим к текстовому вектору две оставшиеся модальности: категориальные признаки и численные признаки из табличных данных путем простой конкатенации. На рис. 2b видно, что добавление метаданных в виде табличных данных дает улучшение на 0.1 FPR, ROC-AUC равен теперь 0.94. Распределения внутриклассой и межклассовой близости изменилось (2e). Новые модальности добавили

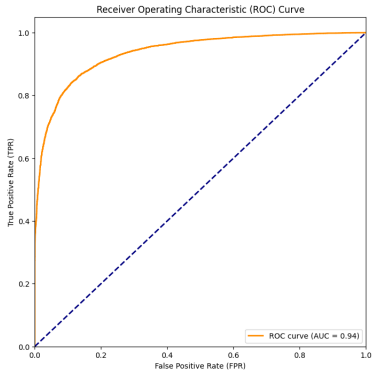
информации, уменьшив хвосты распределения. Однако уверенность в отнесении объекта к своему классу снизилась, потому что простая конкатенация не проецирует признаки разных модальностей в общее пространство и не учитывает связи и зависимости между модальностями.

### 4.3 Добавление модальностей: фреймворк

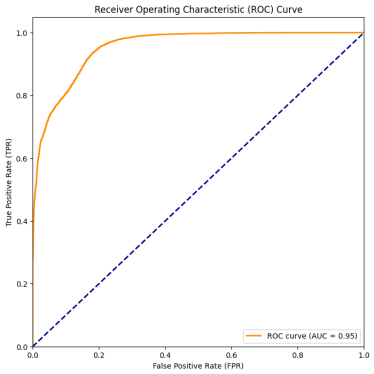
Добавим к текстовому вектору две оставшиеся модальности: категориальные признаки и численные признаки из табличных данных и обучим общее векторное представление с помощью предложенной модели. Обучили 10 эпох, с размером батча равным 512. Каждая эпоха составила 142 итерации и заняла в среднем 4 мин. 30 сек. ROC-AUC вырос на 0.2 относительно текстового трансформерного бейзлайна и составил 0.95 (2c). Распределения внутриклассовой и межклассовой косинусно близости сильно изменились (2f): фреймворк сильнее раздвинул разные категории: объекты из разных классов стали еще более непохожи, в то время как близость объектов внутриклассов увеличилась.



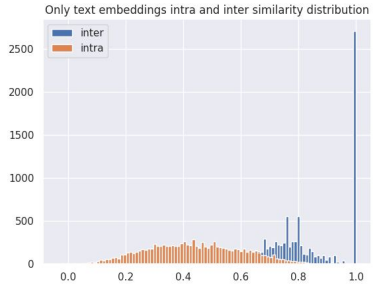

(a) ROC-AUC = 0.93 на векторных представлениях, построенных только на текстовой модальности



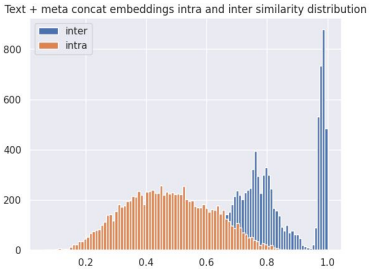

(b) ROC-AUC = 0.94 на векторных представлениях, построенных на 3 модальностях



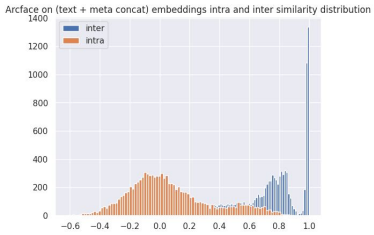

(c) ROC-AUC = 0.95 на обученных векторных представлениях, построенных на 3 модальностях




(d) Внутриклассовая и межклассовая близость векторных представлений, построенных только на текстовой модальности




(e) Внутриклассовая и межклассовая близость векторных представлений, построенных на 3 модальностях




(f) Внутриклассовая и межклассовая близость обученных векторных представлений, построенных на 3 модальностях

Рис. 2: Результаты экпериментов

Таким образом, базовые эксперименты подтверждают, что предложенная идея использовать сферическую классификационную голову и аддитивную угловую функцию потерь для обучения векторных представлений мультимодальных данных работает. В базовой конфигурации фреймворк позволил быстро обучить разделимые векторные представления на трех модальностях: текст, категориальные и табличные данные, увелив ROC-AUC до 0.95.

## 5 Заключение

В работе был предложен новый фреймвок для обучения векторных представлений на мультимодальных данных. Архитектура фреймворка состоит из двух основных компонентов: 1) двухуровневой архи-

тектуры на базе трансформеров для смешения признаков разных модальностей и 2) сферической классификационной головы с аддитивной угловой функцией потерь, напрямую оптимизирующей сферическое расстояние между объектами. Эксперименты показали, что данный подход даже в базовой конфигурации позволяет ускорить обучение и получить более разделимые, геометрически интерпретируемые векторные представления в сравнении с классическими подходами. Предложенная архитектура легковесна, эффективно скалируется при увеличении количества данных и модальностей, не требуя перебора комбинаторного числа пар, как контрастивное обучение. Данная работа открывает новый горизонт для исследований в области обучения мультимодальных векторных представлений.

## 6 Дальнейшие исследования

В работе была изложена только базовая архитектура и проведены простые эксперименты на небольшом датасете, в дальнейшем планируется усложнить архитектуру смешивающей сети, добавив, например, внутренние и кросс-модальные слои внимания, проэксперементировать с добавлением новых модальностей, сравнить результат применения предложенной модели с SOTA решениями на бенчмарк-датасетах, протестировать обученные векторные представления на реальных доменных задачах.

## Список литературы

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 14(3):478–493, 2020a. doi:10.1109/JSTSP.2020.2987728.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 689–696, 2011.

Xuehan Bai, Yan Li, Yanhua Cheng, Wenjie Yang, Quan Chen, and Han Li. Cross-domain product representation learning for rich-content e-commerce. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 5697–5706, October 2023.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3942–3951, 2021.

Md Mofijul Islam and Tariq Iqbal. Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition. IEEE Robotics and Automation Letters, 6(2):1729–1736, 2021. doi:10.1109/LRA.2021.3059624.

Su-Fang Zhang, Jun-Hai Zhai, Bo-Jun Xie, Yan Zhan, and Xin Wang. Multimodal representation learning: Advances, trends and challenges. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC), pages 1–6, 2019. doi:10.1109/ICMLC48188.2019.8949228.

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 14(3):478–493, 2020b.

Narjes Nikzad-Khasmakhi, Mohammad Ali Balafar, M Reza Feizi-Derakhshi, and Cina Motamed. Berters: Multimodal representation learning for expert recommendation system with transformers and graph embeddings. Chaos, Solitons & Fractals, 151:111260, 2021.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. IEEE Access, 7:63373–63394, 2019. doi:10.1109/ACCESS.2019.2916887.

Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C. Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 21252–21262, June 2022.

Zhenhua Huang, Xin Xu, Juan Ni, Honghao Zhu, and Cheng Wang. Multimodal representation learning for recommendation in internet of things. IEEE Internet of Things Journal, 6(6):10675–10685, 2019.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4690–4699, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. arXiv preprint arXiv:2005.08314, 2020.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. Tabbie: Pretrained representations of tabular data. arXiv preprint arXiv:2105.02584, 2021.

Jiacheng Li, Jingbo Shang, and Julian McAuley. Uctopic: Unsupervised contrastive learning for phrase representations and topic mining. arXiv preprint arXiv:2202.13469, 2022.

An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. Personalized showcases: Generating multi-modal explanations for recommendations. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2251–2255, 2023.