Создание мультимодальной модели на основе нескольких визуальных энкодеров

Прокопов Егор*, Тисленко Максим*, Сурков Егор†
Золотарев Дмитрий*, Крюков Андрей*

1. Введение

В современных исследованиях по обработке и анализу данных активно используются различные модальности, что требует применения специализированных энкодеров для извлечения представлений. Однако эффективность тех или иных энкодеров может существенно варьироваться в зависимости от конкретной задачи. В связи с этим, перспективным направлением становится использование процедуры смешивания представлений, полученных из различных энкодеров.

В данной работе мы ставим перед собой следующие задачи:

Достижение уровня state-of-the-art (SOTA) в решении выбранной задачи.

Исследование и тестирование различных методов смешивания эмбеддингов, полученных из различных энкодеров.

Оценка эффективности различных подходов к комбинированию энкодеров с целью улучшения качества представлений.

Таким образом, целью работы является оптимизация процесса получения представлений и их интеграции, что может привести к значительному улучшению результатов в задачах обработки данных.

2. Существующее решение в OmniFusion

Левая часть изображения описывает стандартное решение, в котором используется один визуальный энкодер для обработки изображений. Вот ключевые моменты, которые можно выделить из схемы:

1) Использование одного визуального энкодера: Визуальный энкодер представлен моделью CLIP-ViT-L. Это означает, что все визуальные данные (например, изображения) проходят через этот единый энкодер для получения эмбеддингов.

2) Интеграция эмбеддингов: Изображения поступают на вход через <IMG_START> токен и обрабатываются через энкодер CLIP-ViT-L. Полученные эмбеддинги изображений (отмечены жёлтым цветом) передаются через адаптер и объединяются с текстовыми эмбеддингами (отмечены синим цветом), которые генерируются с использованием модели токенизации BPE (Byte Pair Encoding).

3) Формирование единого представления: После объединения эмбеддингов, данные поступают в модель Mistral-7B, которая отвечает за обработку и генерацию ответа на основе совокупности текстовых и визуальных данных. Цель

*Университет ИТМО, Санкт-Петербург
†Тульский государственный университет, Тула

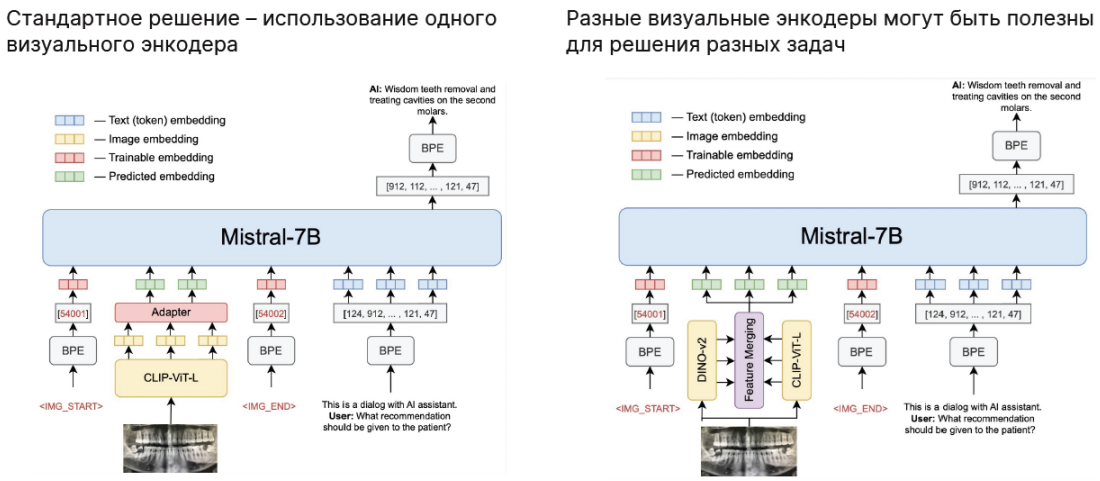

Рис. 1. Существующее решение и возможное улучшение

модели: Система демонстрирует задачу, в которой ассистент на основе анализа визуальной информации (например, рентгеновского снимка зубов) и текстового запроса генерирует ответ. Таким образом, существующее решение подразумевает использование единственного визуального энкодера для обработки всех изображений, что упрощает архитектуру, но может ограничивать её гибкость и адаптивность при работе с разными типами визуальных данных.

Правая часть изображения иллюстрирует более сложное решение, в котором используется несколько визуальных энкодеров, что позволяет учитывать особенности различных типов визуальных данных. Вот основные моменты:

1) Использование нескольких визуальных энкодеров: В отличие от левой части, здесь применяются два разных визуальных энкодера: DINO-v2 и CLIP-ViT-L. Это позволяет системе учитывать специфику различных задач или типов изображений, используя наиболее подходящий энкодер для каждого случая.

2) Объединение признаков (Feature Merging): Эмбеддинги, полученные от двух визуальных энкодеров, проходят через процедуру объединения признаков (Feature Merging). Это позволяет совместить информацию, полученную из разных источников, для создания более богатого представления.

3) Интеграция эмбеддингов: Как и в стандартном решении, полученные эмбеддинги изображений объединяются с текстовыми эмбеддингами, которые генерируются с использованием модели BPE. Данные снова поступают в модель Mistral-7B для дальнейшей обработки и генерации ответа.

4) Гибкость и адаптивность: Благодаря использованию нескольких визуальных энкодеров, система становится более гибкой и способной эффективно обрабатывать различные типы изображений, что может быть полезно для решения разнообразных задач. Таким образом, расширенное решение в правой части изображения демонстрирует подход, в котором разные визуальные энкодеры могут быть использованы для решения различных задач. Это повышает

адаптивность модели и позволяет ей лучше справляться с различными типами визуальных данных, объединяя их для создания более точных и информативных представлений.

3. Возможные решения

На слайде представлены различные варианты смешивания эмбеддингов и описание метода из работы "Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models". Рассмотрим возможные решения на основе представленного материала.

Варианты смешивания эмбеддингов

- Аддитивное смешивание

$$z = w_1 x_A + w_2 x_B$$

  В этом методе смешивание эмбеддингов осуществляется путем взвешенного суммирования. Веса w_1 и w_2 определяют вклад каждого эмбеддинга (например, от разных энкодеров или модальностей) в итоговое представление z. Аддитивное смешивание позволяет легко интегрировать информацию из различных источников и использовать её для дальнейших вычислений. Это простой и эффективный метод для объединения разнородных данных.
- Мультипликативное смешивание

$$z = w_1(x_A \times x_B)$$

  В данном методе смешивание выполняется через взвешенное произведение эмбеддингов. Это может быть полезно для улавливания более сложных зависимостей между различными модальностями. Мультипликативное смешивание может быть более информативным в тех случаях, когда между эмбеддингами существует сильная корреляция или взаимодействие.

Пример решения из работы "Vary"

В этой работе описан процесс тренировки небольшой языковой модели (LLM) с визуальным словарем на основе OPT с 125 млн параметров. Сначала обрабатываются изображения из документов и других источников, таких как ручные заметки или фотографии. Затем эмбеддинги, полученные с помощью модели CLIP и визуально обученного словаря, конкатенируются и подаются в языковую модель размером 7 млн параметров для генерации текста.

Основные элементы метода:

- The new vocabulary network: новая сеть, созданная для обработки визуальной информации и создания эмбеддингов.
- Input embedding from the large language model: используется как базовое представление для генерации текста.

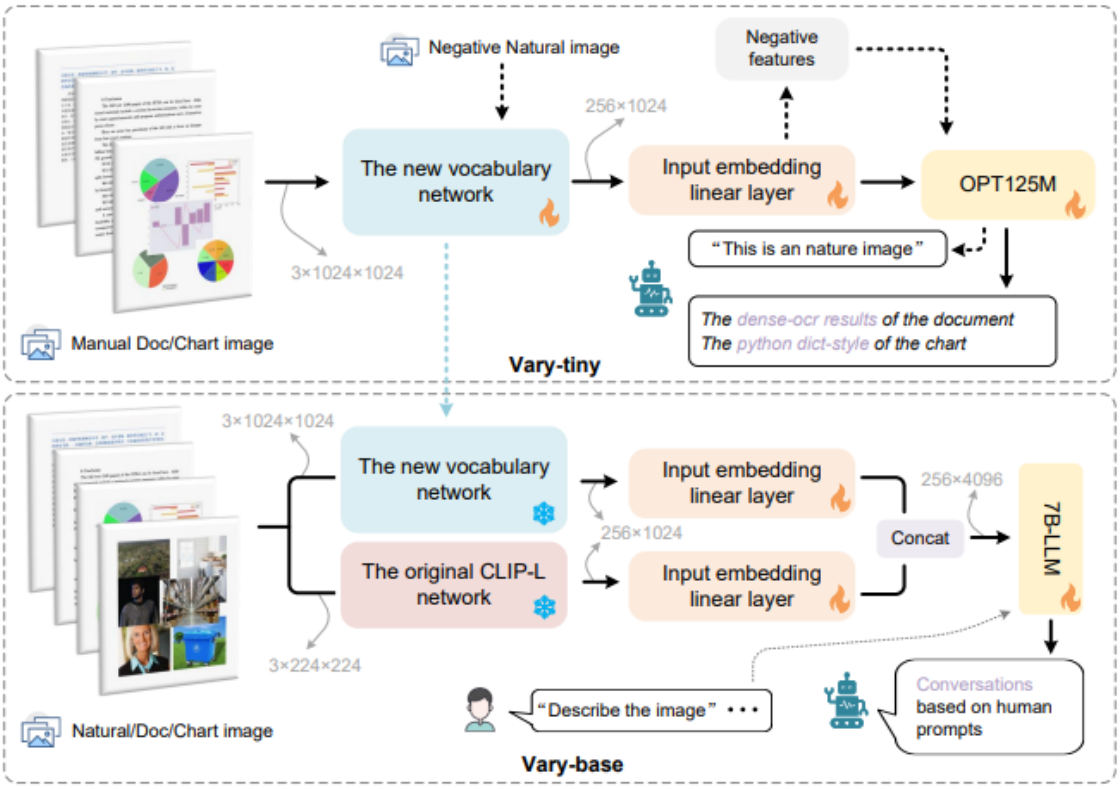

Рис. 2. Решение, предложенное Vary

- Concatenation: объединение эмбеддингов, полученных с помощью CLIP и визуального словаря, перед их подачей в LLM.

Эти подходы могут помочь в создании более мощных и гибких моделей, способных эффективно работать с разными типами данных и задачами.

4. Работа и результаты

- Найденные наборы данных:
  - OCR (Optical Character Recognition): Найдены и проанализированы наборы данных для задач оптического распознавания символов. Эти данные используются для обучения и тестирования моделей, которые могут распознавать текст на изображениях.
  - LaTeX: Собраны данные для обработки и генерации документов в формате LaTeX. Это может включать как тексты, так и различные элементы форматирования.
  - Object Detection: Найдены наборы данных, предназначенные для задачи обнаружения объектов на изображениях. Эти данные содержат изображения с аннотациями, указывающими на расположение объектов.
  - Plots (Графики): Собраны данные для обработки графиков, что может включать изображения графиков с различными типами визуализаций.

  - Tables (Таблицы): Найдены наборы данных, содержащие таблицы, которые используются для обучения моделей на задачи извлечения и обработки информации в табличном формате.
- Обучение ldp адаптера для задачи Object Detection:
  - Проведено обучение адаптера для задачи обнаружения объектов (OD) на основе модели Codetr. Адаптеры помогают в интеграции и улучшении существующих моделей под специфические задачи, такие как Object Detection, путем использования предварительно обученных моделей и их настройки на новые задачи.
- Обучение адаптера CLIP для модели qwen-0.5B:
  - Обучен адаптер CLIP, который используется для модели qwen-0.5B. CLIP (Contrastive Language-Image Pre-training) позволяет модели лучше справляться с задачами, связанными с сопоставлением текста и изображений, улучшая результаты на таких задачах как генерация описаний и поиск.
- Создание датасета на базе COCO для задачи image captioning:
  - Создан новый датасет на базе COCO (Common Objects in Context) для задачи генерации описаний изображений (image captioning). Этот датасет направлен на улучшение понимания расположения объектов на изображениях, что позволяет моделям более точно генерировать текстовые описания.
- Обучение адаптера для LaTeX:
  - Проведено обучение адаптера, который предназначен для обработки и генерации документов в формате LaTeX. Это может включать адаптацию модели для работы с LaTeX-форматированными документами и улучшение качества их генерации.
- Разработка кода для датасетов и адаптеров:
  - Написан код для работы с датасетами и адаптерами, относящимися к изображению графиков и таблиц. Этот код включает в себя подготовку данных, их преобразование и обучение адаптеров, что позволяет эффективно использовать данные для последующего анализа и обработки.

5. Планы

- Протестировать различные методы комбинирования энкодеров:
- Попробовать различные методы смешивания эмбеддингов:
  - Будут исследованы различные подходы к смешиванию эмбеддингов, полученных от различных источников или моделей. Это может включать методы линейного и нелинейного смешивания, а также использование обучаемых слоев для комбинирования эмбеддингов. Целью является улучшение представлений данных и повышение производительности моделей.
- Добиться достижения состояния SOTA (State-of-the-Art):
  - Основной целью является достижение состояния SOTA для выбранных задач и методов. Это потребует тщательного подбора гиперпараметров, оптимизации архитектур моделей и проведения

многочисленных экспериментов. Будут применяться современные методы и подходы, чтобы обеспечить конкурентоспособность результатов.

- Разработать и протестировать метод, аналогичный тому, что приведен в работе “Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models”:
  - Исследование метода, описанного в статье “Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models”, и разработка аналогичной методологии для улучшения масштабируемости словаря визуальных данных в больших моделях. Планируется тестирование разработанного метода и оценка его эффективности по сравнению с существующими подходами.

6. Итоги

В результате было обучено решение для получения представлений и была заготована интеграция в существующее решение OmniFusion.

7. Благодарности

Эта работа не могла бы быть сделана без помощи наших менторов: Матвея Скрипкина, который всегда готов прийти на помощь и разбираться в нашем коде, и Елизаветы Гончаровой, которая может великолепно объяснить самые сложные моменты, они классные)

8. Литература

[1] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, Xiangyu Zhang, Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models, arXiv preprint arXiv:2312.06109 (2023)

[2] Elizaveta Goncharova, Anton Razzhigaev, Matvey Mikhalchuk, Maxim Kurkin, Irina Abdullaeva, Matvey Skripkin, Ivan Oseledets, Denis Dimitrov, Andrey Kuznetsov, OmniFusion Technical Report, arXiv preprint arXiv:2404.06212 (2024)