

ДВУХФАЗНАЯ ШЛЮЗОВАЯ ИНТЕГРАЦИЯ ТЕКСТОВЫХ, ВИЗУАЛЬНЫХ И ТАБЛИЧНЫХ ДАННЫХ В ЯЗЫКОВОЙ МОДЕЛИ

ГАПОНОВ Н. П., гр. ИИБ-231, III курс

Кузбасский государственный технический университет имени Т.Ф. Горбачева
г. Кемерово

Интеграция разнородных данных в языковой модели требуется в задачах, где текстовый запрос сопровождается изображением, таблицей измерений или структурированной записью. Простое присоединение всех признаков к текстовой последовательности не учитывает различие масштабов, семантики и надёжности модальностей. Визуальный поток состоит из большого числа локальных фрагментов, табличный – из небольшого набора признаков с разными единицами измерения, а текстовый поток задаёт синтаксический контекст и порядок вывода. Поэтому модель должна получать совместное представление, не разрушая внутреннюю структуру каждого источника и не увеличивая вычислительную стоимость пропорционально сумме всех токенов.

Базовый механизм внимания сопоставляет пары элементов последовательности и формирует контекстное представление с помощью матриц запросов, ключей и значений [1]. Для изображения такой подход естественно сочетается с разбиением на фиксированные патчи, используемым в Vision Transformer [2]. Таблица не имеет единственного пространственного порядка, поэтому её строки и столбцы сначала преобразуются в набор признаков с явным кодированием типа поля, единицы измерения и признака пропуска. Общая задача первой фазы состоит в том, чтобы привести выходы трёх энкодеров к одному скрытому размеру d , сохранив маску происхождения каждого признака.

В качестве семантического ориентира для первой фазы удобно использовать контрастивное согласование представлений: близкие по смыслу текстовое описание и изображение притягиваются, а несоответствующие пары разводятся [3]. Для таблиц вместо изображения применяются пары «описание – запись» и проверка согласованности числовых полей. Важно, что согласование не требует сливать исходные токены. Каждый энкодер продолжает работать в собственной координатной системе, а обучаемый линейный адаптер переводит результат в пространство языковой модели.

Проблема второй фазы заключается в размере совместного контура внимания. Если T , V и U – числа текстовых, визуальных и табличных токенов, то полная конкатенация требует вычислить $(T + V + U)^2$ парных взаимодействий. При увеличении разрешения изображения величина V быстро становится

5.21.2

доминирующей, хотя языковой модели для формулировки ответа обычно достаточно компактного набора визуальных свидетельств. Поэтому в предлагаемой схеме используется фиксированное число m шлюзовых query-токенов, которые извлекают информацию из модальностей перекрёстным вниманием. Идея информационного бутылочного горлышка соответствует архитектуре Perceiver [4].

Схема интеграции показана на рисунке 1. На первой фазе для каждой модальности выполняются маскирование пропусков, нормировка и проекция в размерность d . На второй фазе m query-токенов обращаются к объединённой последовательности адаптеров, после чего полученные токены вводятся в несколько слоёв языковой модели через остаточную ветвь. Визуальные и табличные признаки не добавляются к каждому текстовому токenu: они проходят через один и тот же компактный интерфейс. Это позволяет сохранить текстовый контур модели и заменить отдельный энкодер без изменения словаря.



Рисунок 1 – Двухфазный контур шлюзовой интеграции мультимодальных данных

Для каждого источника x_m адаптер A_m формирует последовательность признаков P_m размерности d . Перед проекцией применяется нормировка, зависящая только от активных элементов модальности; нулевая маска не должна порождать искусственный сигнал. Формально получаем

$$p_m = A_m(LN(x_m)) \quad (1)$$

Здесь LN обозначает нормировку по скрытому размеру, а индекс m принимает значения *text*, *vision* и *table*. Вектор маски хранится отдельно и передаётся в шлюз, поэтому отсутствие изображения отличается от изображения, состоящего из почти нулевых пикселей. Для числовых столбцов до LN выполняется робастное масштабирование по медиане и межквартильному размаху; категориальные поля получают обучаемое кодирование. Такие операции

уменьшают влияние единиц измерения и предотвращают доминирование одного столбца над остальными.

Шлюзовые токены Q взаимодействуют с последовательностью $P = [P_{\text{text}}; P_{\text{vision}}; P_{\text{table}}]$. Полученное состояние нормируется и добавляется к скрытому состоянию h языковой модели только после умножения на обучаемый коэффициент g_m . В простейшем варианте коэффициент является скаляром для модальности, в расширенном – вектором по каналам. Обновление слоя записывается как

$$h' = LN(h + \sum_m g_m P_m Q) \quad (2)$$

Функция g_m ограничивается интервалом $[0, 1]$ сигмоидой и вычисляется из агрегированного состояния модальности. Она выполняет роль проверяемого предохранителя: при отсутствии источника или конфликте с текстовым контекстом вклад соответствующей ветви уменьшается, а не маскируется постфактум в выходном ответе. В обучении полезно применять случайное отключение модальностей и штраф за чрезмерное насыщение всех гейтов единицей. Тогда модель учится формулировать ответ по доступным источникам и не использует пустой канал как скрытый идентификатор.

После шлюза в языковой модели присутствуют только T текстовых и m мультимодальных токенов. Для примера рассмотрим $T = 128$, $V = 196$, $U = 12$ и $m = 16$. При полной конкатенации длина контура равна 336, а число парных взаимодействий – 112 896. В шлюзовой схеме текстовое самовнимание требует 128^2 взаимодействий, а извлечение свидетельств – $16 \cdot (196 + 12)$. Суммарная оценка составляет 19 712, то есть 17,5 % от полной стоимости. Линейные операции адаптеров и нормировки в эту оценку не входят, поскольку они растут пропорционально числу токенов и не образуют квадратичную матрицу.

Таблица 1 – Сравнение вычислительной стоимости интеграционных схем

Схема	Токены в совместном контуре	Парные взаимодействия	Относительная стоимость
Полная конкатенация	$T + V + U = 336$	112 896	100 %
Шлюзовая интеграция	$T + m = 144$	19 712	17,5 %
Только текст	$T = 128$	16 384	14,5 %

Сокращение длины совместного контура не означает отказ от визуальной детализации. Все V патчей участвуют в перекрёстном внимании шлюза, но только m агрегированных состояний передаются в языковой контур. Это близко по идее к визуально-языковой модели Flamingo, где между замороженными компонентами вводятся специальные слои перекрёстного внимания [5]. В BLIP-2 аналогичный интерфейс Q-формера связывает замороженный визуальный энкодер и большую языковую модель [6]. Отличие предлагаемого варианта состоит в симметричном отношении к таблице: для неё предусмотрен такой же адаптер и такой же канал маски, а не отдельная ветвь постобработки.

5.21.4

Чтобы не переобучать всю языковую модель, параметры можно добавлять в виде низкоранговых поправок к матрицам шлюза и выбранным проекционным слоям. Метод LoRA показывает, что обновление малой добавочной матрицы сохраняет базовые веса и уменьшает объём оптимизируемых параметров [7]. В данной постановке это особенно удобно при последовательном подключении новых типов таблиц или датчиков: меняются адаптер и небольшая часть шлюзовых слоёв, тогда как языковое ядро остаётся общим для различных наборов модальностей.

Практический протокол обучения содержит три одновременно действующих слагаемых. Авторегрессионная ошибка по целевому тексту проверяет, что интеграция помогает решать задачу. Контрастивное слагаемое выравнивает описание с визуальным и табличным содержанием. Слагаемое согласованности сравнивает ответы при исходном наборе модальностей и при случайном отключении одного источника. Последнее не должно заставлять ответы совпадать буквально; его назначение – ограничить необоснованное изменение фактов, которые присутствуют в оставшихся каналах. При наличии инструкций пользователя можно дополнительно использовать мультимодальное instruction-tuning [8], но маски и гейты при этом сохраняются.

Корректность интеграционного контура проверяется на трёх классах случаев. В полном случае все модальности имеют ненулевую маску, размерности адаптеров совпадают, а сумма гейтов находится в допустимом диапазоне. В случае отсутствия изображения визуальная маска обнуляется до шлюза; выход сохраняет конечные значения и зависит только от текста и таблицы. В случае противоречивого числового признака проверяется, что нормировка выполняется до проекции, поэтому изменение единиц измерения не меняет размерность скрытого состояния. Эти проверки относятся к вычислительному контракту и выполняются до оценки качества генерации.

Для приведённого примера расчёт можно записать общей формулой. Если C – число каналов, которое реально подаётся в языковую модель, а d – её скрытый размер, то стоимость квадратичной части контура равна

$$C = T^2 + m(V + U) + m^2 \quad (3)$$

Слагаемое m^2 учитывает внутреннее взаимодействие шлюзовых токенов. При $m = 16$ оно равно 256 и мало по сравнению с T^2 . Поэтому увеличение числа query-токенов следует связывать не с числом исходных патчей, а с требуемой сложностью задачи: для мелких объектов и точных числовых зависимостей m можно увеличить, не возвращаясь к полной конкатенации. Одновременно размер m должен быть зафиксирован в конфигурации модели, чтобы длина контекста и расход памяти были известны до запуска.

В рассматриваемом вычислительном примере при $T = 128$, $V = 196$, $U = 12$ и $m = 16$ число элементов, участвующих в формировании матрицы внимания, составляет 112 896 для полной конкатенации, 19 712 для шлюзового варианта и 16 384 для текстовой схемы. Таким образом, отношение

вычислительной нагрузки шлюзовой схемы к полной составляет $19\,712 / 112\,896 = 0,1746$, или около 17,5 %. Полученное соотношение характеризует исключительно объём квадратичной части вычислений. Само по себе уменьшение числа взаимодействий не позволяет сделать вывод о сохранении точности, поэтому сравнение качества необходимо проводить отдельно для случаев с полным набором входов, отсутствующими модальностями и противоречивыми данными.

Сжатие исходного набора признаков до фиксированного числа m состояний сопровождается риском утраты локальной информации. Это особенно существенно в ситуациях, когда требуется извлечь из изображения конкретный пространственный фрагмент либо выполнить расчёт по отдельным элементам протяжённой таблицы. Для таких задач целесообразно использовать каскадный вариант шлюза. На первом этапе он формирует компактное представление содержания модальностей, после чего по запросу языковой части выполняется дополнительное обращение к выбранным исходным признакам. Такой механизм увеличивает время обработки, однако позволяет сохранить ограниченную длину основного контекста и одновременно обеспечить трассируемость информации, использованной при формировании ответа.

Отдельное ограничение связано с необходимостью согласованного представления данных различных типов. Контрастивное выравнивание обеспечивает близость связанных представлений, однако само по себе не контролирует корректность единиц измерения, временных характеристик и пространственной привязки изображения. Поэтому при обработке таблиц вместе со значением признака следует сохранять его тип, единицу измерения и информацию о пропуске. Для визуальных данных дополнительно целесообразно учитывать разрешение и заданную область анализа. Если источники содержат противоречивые сведения, в результирующем представлении должна сохраняться информация о том, какая модальность использовалась. В задачах, предъявляющих повышенные требования к достоверности, этот механизм следует дополнять специализированной проверкой фактических и числовых результатов.

Разделение интеграционного контура на две фазы позволяет изменять способы подключения данных без перестройки языкового компонента. Подключение нового типа входной информации сводится к разработке соответствующего адаптера и настройке процедуры нормировки. Параметры второй фазы могут изменяться независимо – например, за счёт выбора количества шлюзовых состояний или положения соответствующих слоёв. При этом словарь и основная текстовая часть языковой модели сохраняются. Для обеспечения воспроизводимости конфигурация эксперимента должна фиксировать значения T , V , U и m , используемые маски, версии адаптеров и контрольные суммы соответствующих наборов весов.

Предложенная схема выполняет интеграцию текстовой, визуальной и табличной информации посредством общего скрытого пространства и ограниченного

набора шлюзовых состояний. В первой фазе входные представления нормируются и приводятся к единой размерности с сохранением сведений о наличии соответствующего источника. Далее шлюз извлекает из них наиболее значимую информацию и передаёт в языковой контур уже компактное представление. Для $T = 128$, $V = 196$, $U = 12$ и $m = 16$ это приводит к снижению квадратичной составляющей вычислительной нагрузки до 17,5 % относительно полной конкатенации. Использование управляемых коэффициентов и масок позволяет учитывать неполный состав входных данных, а модульная организация адаптеров – расширять перечень поддерживаемых форматов без изменения языкового ядра. Для задач, в которых критична детализация исходной информации, шлюз может дополняться механизмом повторного обращения к исходным признакам и отдельным контуром верификации.

Список литературы:

1. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.
2. Dosovitskiy A., Beyer L., Kolesnikov A. et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale // International Conference on Learning Representations. – 2021. – URL: <https://openreview.net/forum?id=YicbFdNTTy> (дата обращения: 31.08.2026).
3. Radford A., Kim J. W., Hallacy C. et al. Learning Transferable Visual Models From Natural Language Supervision // Proceedings of the 38th International Conference on Machine Learning. – 2021. – Vol. 139. – P. 8748–8763.
4. Jaegle A., Borgeaud S., Alayrac J.-B. et al. Perceiver: General Perception with Iterative Attention // Proceedings of the 38th International Conference on Machine Learning. – 2021. – Vol. 139. – P. 4651–4664.
5. Alayrac J.-B., Donahue J., Luc P. et al. Flamingo: a Visual Language Model for Few-Shot Learning // Advances in Neural Information Processing Systems. – 2022. – Vol. 35. – P. 23716–23736.
6. Li J., Li D., Savarese S., Hoi S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models // Proceedings of the 40th International Conference on Machine Learning. – 2023. – Vol. 202. – P. 19730–19742.
7. Hu E. J., Shen Y., Wallis P. et al. LoRA: Low-Rank Adaptation of Large Language Models // International Conference on Learning Representations. – 2022. – URL: <https://openreview.net/forum?id=nZeVKeeFYf9> (дата обращения: 31.08.2026).
8. Liu H., Li C., Wu Q., Lee Y. J. Visual Instruction Tuning // Advances in Neural Information Processing Systems. – 2023. – Vol. 36. – URL: <https://arxiv.org/abs/2304.08485> (дата обращения: 31.08.2026).