Моделирование и анализ данных
2024. Том 14. № 4. С. 45–62
DOI: https://doi.org/10.17759/mda.2024140403
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2024. Vol. 14, no. 4, pp. 45–62
DOI: https://doi.org/10.17759/mda.2024140403
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.89

Метод синтеза поведения когнитивного агента на основе обработки мультимодальных сигналов

Вейценфельд Д.А.*
Российский Университет Дружбы Народов им. Патриса Лумумбы (РУДН)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-2787-0714
e-mail: veicenfeld@isa.ru

Киселёв Г.А.**
Федеральный Исследовательский Центр
«Информатика и Управление» Российской Академии Наук (ФИЦ ИУ РАН)
г. Москва, Российская Федерация
Российский Университет Дружбы Народов им. Патриса Лумумбы (РУДН)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-9231-8662
e-mail: kiselev@isa.ru

В статье рассматривается проблема прогнозирования деятельности агента исходя из текстового описания задачи и визуального анализа среды. Предложено обновление подходов классической когнитивной архитектуры, позволяющее применять её в реальной среде. Разработано дополнение семиотического метода символьного обозначения авторским нейросетевым механизмом связывания векторов текстового и визуального пространств. Проведен ряд экспериментов с полученной моделью в комплексной среде эмулятора вождения автомобиля.

Ключевые слова: Планирование поведения, обучение с подкреплением, когнитивные архитектуры, оценка ситуации, мультимодальный анализ.

Благодарности. Публикация выполнена при поддержке Программы стратегического академического лидерства РУДН, проект № 021934-0-000.
Работа выполнялась с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

Для цитаты:
Вейценфельд Д.А., Киселев Г.А. Метод создания поведения когнитивных агентов на основе обработки мультимодальных сигналов // Моделирование и анализ данных. 2024. Том 14. № 4. С. 45–62. DOI: https://doi.org/10.17759/mda.2024140403

***Вейценфельд Даниил Анатольевич,** студент магистратуры, Российский Университет Дружбы Народов им. Патриса Лумумбы (РУДН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-2787-0714, e-mail: veicenfeld@isa.ru

****Киселев Глеб Андреевич,** кандидат технических наук, научный сотрудник, Федеральный Исследовательский Центр «Информатика и Управление» Российская Академия Наук (ФИЦ ИУ РАН); старший преподаватель, Российский Университет Дружбы Народов им. Патриса Лумумбы (РУДН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-9231-8662, e-mail: kiselev@isa.ru

1. ВВЕДЕНИЕ

Когнитивные архитектуры предназначены для повышения автономности робототехнических платформ и позволяют им выполнять сложные задачи без человеческого вмешательства. Архитектура должна обеспечивать сложное поведение агента, при наличии группы, осуществлять внутригрупповую коммуникацию и оптимизировать деятельность агентов [1–6]. Интеллектуальный подход к разработке архитектур является актуальной задачей и требует решения множества междисциплинарных проблем. Обычно, структура когнитивных архитектур делится на три уровня контроля поведения: стратегический, тактический и реактивный [7]. Эти уровни реализованы в одной из наиболее перспективных архитектур на сегодняшний день [8], в которой продемонстрированы особенности взаимодействия методов интеллектуального управления, планирования траекторий и поведения. Однако, в предыдущих исследованиях, посвященных архитектуре, недостаточно внимания уделялось реализации функций межмодальных взаимосвязей, представляя этот механизм в виде абстрактных математических переходов между типами данных.

Цель данной работы заключается в предложении методов программной реализации функций-переходов для межмодальных взаимосвязей системы на основе современных методов представления знаний. Подцелями являются описание реализации функции создания последовательности целенаправленных действий когнитивного агента с описываемой архитектурой и ознакомление с методами решения ряда технических проблем, возникающих при реализации действий на робототехническом агенте.

Большинство современных когнитивных архитектур имеет модульную иерархическую структуру [9], где каждый уровень отвечает за выполнение определенных робототехнических функций. В настоящей работе представлен способ верхнеуровневого планирования действий, распознавание команд на естественном языке и анализ изображения с камер агента. Решается задача поиска оптимального поведения на основе сформулированной на естественном языке цели в неизвестной заранее агенту карте. Основным отличием от общеизвестных методов построения управления на основе известной функции переходов между состояниями в подходе является предположение о возможности синтеза этой функции с помощью встроенных нейросетевых методов внимания, которые влияют на динамическое планирование поведения агента.

Когнитивная составляющая предлагаемой архитектуры ранее была основана на семиотическом представлении знаний [10], базовым элементом которой является

знак [11]. Использование знаковой парадигмы в качестве способа представления знаний об элементах окружающей среды было выбрано исходя из исследований нейрофизиологов и психологов [12, 13] относительно представления знаний человеком. Знаковая парадигма архитектуры учитывает различные типы информации об одном и том же объекте и использует их для перехода от классического для искусственного интеллекта символьного способа представления знаний к гибридному, который отвечает принципам робототехники настоящего времени. В рассматриваемом подходе, явное представление функции семантического связывания знаний агента представлена нечетким образом.

Знак описывается кортежем, в который включаются 4 основные компоненты $s = \langle n, p, m, a \rangle$, где n – компонента имени, с помощью которого задаётся знак, p – компонента образа, m – компонента значения и a – компонента смысла. Каждая из компонент знака отвечает за свой вид представления информации об описываемой сущности. Компонента p является описанием характерных признаков сущности (при геометрической постановке задачи признаки сущности дополняются описанием псевдофизических представлений окружающего пространства [8, 14]). Компонента m – отвечает за доступные обобщенные сценарии использования сущности коллективом агентов, а компонента a – определяет роль представленной знаком сущности в действии планирующего субъекта. Смыслы знака синтезируются в процессе деятельности агента с описываемой сущностью и являются актуализацией в рамках настоящей задачи значений этого знака. Знаком может быть представлен как статический объект, так и действие.

Компоненты знаков связаны и формируют семантические сети, отношения на каждой из которых различны (например, на сети m – объект-роль, на сети p – часть-целое, а на сети a – коалиция – участник коалиции). Робототехнические реализации алгоритмов [8, 9] используют особые предикативные способы представления как описания окружающей агента ситуации, так и его возможных действий и синтезируют последовательность целевых действий агента с помощью алгоритмов распространения активности – последовательного уточнения ролевого состава действий за счёт анализа участников текущей и целевых ситуаций, а также их возможностей.

Описанный способ представления окружающей среды хоть и позволяет привести к общему виду психологически правдоподобные требования к синтезу деятельности робототехнических платформ, но громоздкость используемых структур не позволяет применять его в реальных, не учебных, примерах с должным уровнем автоматизации. Прежде всего, такие задачи возникают при создании систем управления в автономных беспилотных транспортных средствах [15–18]. Основным исключающим фактором является скорость принятия решений когнитивным циклом архитектуры, включающим последовательность рассуждений на каждом из уровней автоматизации. В настоящей статье предложен метод, позволяющий ускорить когнитивный цикл архитектуры и протестировать её в реальных сценариях. Далее статья организована следующим образом. В разделе 1 приведено описание класса решаемых задач и представлен общий алгоритм синтеза поведения с помощью мультимодального

анализа данных. В разделе 2 описана предлагаемая архитектура системы, методика тестирования её работоспособности на современном мультимодальном наборе данных. В разделе 3 приведены результаты экспериментов на сервере ФИЦ ИУ РАН.

2. ПОСТАНОВКА ПРОБЛЕМЫ

Одной из наиболее трудозатратных задач для вычислительного модуля робототехнической реализации когнитивного агента является формирования целенаправленного поведения в виде последовательности иерархических действий, требуемых для решения стоящей перед агентом задачи. Этот процесс состоит из работы 2 основных типов алгоритмов: алгоритма целеуказания и алгоритма синтеза плана поведения. В статье описывается случай, в котором целеуказание производится извне, а агент должен сформировать самостоятельно целевую ситуацию и составить план действий, по её достижению. В рамках настоящей статьи поведение агента и последовательность его действий являются синонимами и далее будут употребляться взаимозаменяющим образом.

В робототехнических платформах план поведения синтезируется набором взаимосвязанных алгоритмов, включающих в себя анализ визуальных и текстовых данных, планирование на их основе, предсказание результатов выполнения действий и других составляющих когнитивного цикла системы. В простейшем случае, в результате работы цикла появляется последовательность абстрактных действий (сценария поведения), каждое из которых имеет уточнение в виде набора моторных команд на сервоприводы робототехнического агента. Для синтеза набора моторных команд используются алгоритмы оптимизации, например, алгоритмы оптимального управления манипулятором и другие, позволяющие реализовывать высокоуровневые команды, не включая их детерминированные составляющие в общий цикл принятия решений. Алгоритмы синтеза оптимального управления можно условно разделить на алгоритмы с дискретным представлением среды – обучение с подкреплением, и алгоритмы непрерывного синтеза управляющего воздействия на основе реакции на сигналы окружающей среды – алгоритмы обработки поведения ПИД-регуляторов. Первый тип методов является развитием теории конечных автоматов и попыткой психологически-правдоподобного решения задачи управления с дискретным временем [19], в нем используются способы изучения среды в терминах спонтанной активности нервной системы, а способы сохранения комплексных действий (опций) представляют, как ассоциативную память агента [20–21]. Второй тип методов сводится к моделированию биологически-правдоподобного анализа окружающей среды мозгом и синтеза управляющего воздействия [22, 23] на среду, с помощью активации имеющихся в арсенале агента атомарных действий. В качестве примера биологически-правдоподобной связи можно привести набор методов по анализу ЭЭГ-сигнала и выявления частотных признаков действий в нем, возникающих как следствие стимула (агенту бросают мяч, и он поднимает руки; агенту показывают видео с движением руки, и он повторяет движение).

При описании задачи в терминах знаковой картины мира реакция агента на внешний стимул описывается функциями $\psi_p^m, \psi_m^a, \psi_a^p$ связывания семантических сетей, которые позволяют выстраивать однозначную взаимосвязь между целевой командой и типом реакции на неё. На основе трудов Г.С. Осипова [24] можно выделить 3 типа «картины мира» агента – рациональная, житейская и мифологическая. Каждый из типов является краевым представлением и естественные агенты комбинируют типы, состовляя свой, гетерогенный тип в силу имеющегося опыта взаимодействия с реальностью. Планирование действий агента в простейшем виде можно описать следующей последовательностью методов:

1. Распознавание команды. Функция перевода из сети имен n в сеть значений m. Основные алгоритмы – диаризация (разделение голосов, отдающих команды), транскрибирование (перевод из аудио в текст), методы анализа речи (различные типы анализа и составление графа зависимостей слов в предложении, понимание смысла фразы и целевого объекта).
2. Создание сценария поведения. В житейской картине мира – выбор сценария из имеющихся. В рациональной картине мира – синтез сценария любым алгоритмом планирования [9]. Процесс целиком происходит на сети значений и основным алгоритмом является перебор на основе удовлетворения ограничений по количеству объектов и агентов в задаче. В настоящем методе этот этап возможно провести в неявном виде с помощью использования нейросетевого подхода. Результатом является синтез гипотезы о правильности применения набора действий для достижения целевой ситуации.
3. Уточнение ролей в сценарии поведения. Функция перехода из сети значений m в сеть образов p. Для выстраивания такого типа связи необходимо взаимнооднозначное соответствие между ролевым представлением объекта в сценарии возможного поведения агента и его ранее встреченным изображением. Основным алгоритмом, позволяющим совершать рассуждения по выбору объекта в каждой роли – VQA [25] или более общий подход VLM [17,18].

Шаги 2–3 для систем, действующих в реальных не распознанных средах требуют более детальной проработки и описаны далее.

VLM как алгоритм выбора оптимального действия агента

В этой статье предлагается метод, который автоматически генерирует функцию вознаграждения для агента на основе текстового описания задачи и распознанных им ранее объектов. Стоящая задача не является новой и в работах [26–27] уже описываются примеры LLM, синтезирующих вознаграждение для Q-функции агента на основе данных среды. В работе [28] используется внутреннее состояние агента для синтеза награды за выполнение действий и текстовое описание результата действия для валидации уровня награды. В настоящей работе любая ситуация описывается в качестве последовательности отношений между объектами в комбинированном свернутом виде VLM. Агент получает целеуказание извне и подбирает наиболее похожую ситуацию в пространстве ситуаций на основе описания (как

в житейской картине мира), а после этого активирует заранее известный сценарий достижения целевой ситуации. Сценарии достижения целевой ситуации известны агенту на базовом уровне, вне контекста текущей локации. Уточнение сценария в поставленной задаче сводится к построению карты местности, локализации агента и выполнению сценария методом поиска возможных локальных действий в контексте текущей карты. Процесс оканчивается сохранением детализированного плана для будущего использования.

Пример:

- Запрос: налей кофе и принеси мне.
- Целевая ситуация: агент с кружкой кофе в манипуляторе около человека.
- Базовый сценарий: агент едет на кухню, нажимает на кнопку на кофе-машине и по получению разрешающего сигнала забирает кружку. Агент вместе с кружкой едет к источнику задачи.
- Измененный сценарий:
  1. Агент формирует кухню как свернутый полносвязным слоем набор описаний (холодильник стоит в комнате, кофеварка стоит на шкафу, плита есть, мойка в комнате) со всеми видами ранее известных ему объектов.
  2. Каждой комнате в доме назначается класс «кухня» – как промежуточной сценарной комнаты.
  3. Агент пытается построить сценарий по достижению целевой ситуации исходя из описания кухни – в 1 из комнат есть мойка, есть стол, но нет кофемашины. В другой комнате есть плита, но нет других объектов – выбирается та комната, вектор описания которой наиболее близок к промежуточной необходимой ситуации.
  4. Достижение необходимой промежуточной ситуации увеличивает награду агента.
  5. После достижения целевой ситуации агент сохраняет последовательность действий как навык (опцию).

Обоснование выбора VLM

Основные компоненты VLM–ViT (Visual Transformers) и LLM (Large Language Model). Выбор комбинации ViT и LLM оправдан необходимостью синтеза мультимодальной информации для построения абстрактных планов поведения. ViT обеспечивает высокое качество интерпретации визуальных данных, тогда как LLM дополняет этот процесс, интегрируя текстовую информацию, знания и находя сложные связи между элементами входных последовательностей.

Базовый механизм VLM в настоящей задаче аналогичен VQA (Visual Question Answering). Опрос модели по ключевым текстовым вопросам позволяет извлечь максимальную пользу от механизмов LLM. Подход позволяет агенту строить планы действий, соответствующие поставленным задачам, с учётом всех аспектов сцены и её описания.

ViT используется для анализа изображений из-за наличия механизма извлечения глобальных и локальных визуальных признаков с высоким уровнем семантической детализации. Механизм внимания ViT предназначен для выявления ключевых

аспектов изображений, что критически важно для реализации задачи перцептивного анализа агентом, где требуется понимание сложных визуальных сцен. Например, ViT позволяет интерпретировать сцены с высокой степенью детализации, учитывать пространственные взаимосвязи объектов и выявлять визуальные шаблоны, необходимые для формирования действий агента.

Комбинирование ViT с LLM (например, GPT[29] или LLaMA[30]) позволяет дополнить обработку изображений пониманием агентом постановки задачи на естественном языке. LLM обеспечивает эффективный анализ текстовых описаний задач и запросов, интерпретацию инструкций, а также возможность учитывать общий контекст, включая ранее накопленные знания о мире. Это позволяет модели генерировать осмысленные последовательности действий, учитывая не только визуальную, но и текстовую информацию.

В задачах, связанных с агентами, которые оперируют в пространстве, важно не только распознавать объекты или отвечать на простые вопросы, но и строить планы на основе сложной визуальной и текстовой информации. Модели ViT-LLM способны идентифицировать объекты и их расположение в пространстве, определять семантические взаимосвязи между элементами сцены, генерировать последовательности действий, соответствующих описанию задачи, таких как перемещение объекта, достижение цели или выполнение комплексного набора операций.

Современные VLM, построенные на основе LLM, таких как GPT, обладают существенными преимуществами по сравнению с традиционными моделями VQA, в которых ответы формируются через задачу классификации (выбор из предустановленного набора ответов). Такие модели ограничены фиксированным набором ответов, что снижает их способность адаптироваться к новым запросам или контекстам. Современные VLM используют генеративный подход, который позволяет формировать уникальные ответы, не ограничиваясь предустановленным списком. Это критически важно в задачах, где требуется высокая степень вариативности или детализации в ответах. Классификационные модели VQA [31–33] неэффективны при работе с вопросами, ответы на которые невозможно заранее предсказать, либо с ситуациями, где требуется детальный ответ. VLM тренируются на масштабных мультимодальных наборах данных, имеют механизм переноса и адаптации ранее полученных знаний на новые задачи, что позволяет существенно расширить базовый набор возможностей системы VQA классификации.

ViT делит изображение на равные по размеру регионы (патчи), которые представляют из себя векторы с информацией об объектах внутри патча и взаимоотношениях с объектами в соседних патчах. Итоговые патчи можно интерпретировать как токены для лингвистической модели и минимизировать фактор особенностей конкретного набора данных. Например, модель может быть обучена на наборе данных, сгенерированном в графическом реалистичном эмуляторе Unreal Engine (например, открытый набор данных DriveLM на основе представления города в симулятора Carla), после чего применяться в реальных условиях.

3. РАЗРАБОТКА ЭКСПЕРИМЕНТАЛЬНОЙ VLM

Основной задачей VLM является согласование представлений, получаемых из изображения, с токенами, используемыми LLM. Сложность решения этой задачи в том, что LLM, такие как GPT, обучены работать с текстовыми токенами из своего предопределённого вокабуляра (словарь токенов в LLM, состоящий из слов и подслов), тогда как нейросеть-анализатор изображений генерирует выходные представления в виде эмбеддингов (векторов), которые изначально находятся в другом объектном пространстве. Основная идея использования ViT заключается в преобразовании изображений в «языковые» токены, которые могли бы напрямую взаимодействовать с предобученной LLM.

ViT представляет изображения через эмбеддинги, которые содержат числовые признаки, отражающие пространственные, цветовые и семантические аспекты изображения. Эти эмбеддинги существенно отличаются от токенов, используемых в LLM, которые представляют текстовые данные как дискретные последовательности из предопределённого словаря. Чтобы обеспечить совместимость, требуется создать механизм преобразования визуальных представлений в текстовые.

Формализация проблемы

Пусть,

$$Z_{ViT} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_\mathbf{N}\}$$

– набор эмбеддингов (представлений), полученных из изображения с помощью предобученной ViT. Здесь каждый $z_i \in R^d$ – это векторное представление одного патча изображения.

$$\mathcal{V}_{LLM} = \{\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_\mathrm{M}\}$$

– множество токенов из вокабуляра LLM, где каждый $t_j \in R^k$.

Проблема заключается в том, что условие $\mathbf{Z}_{ViT} \subset \mathcal{V}_{LLM}$ не выполняется, т. е. пространство эмбеддингов ViT и пространство токенов LLM не пересекаются:

$$\forall \mathrm{z_i} \in \mathrm{Z_{ViT}}, \exists \mathrm{t_j} \in \mathcal{V}_{LLM} : \mathrm{z_i} \neq \mathrm{t_j}$$

Эта несовместимость мешает LLM напрямую обрабатывать выходы ViT, так как z_i не интерпретируются в терминах предобученного вокабуляра LLM.

Использование проекционной прослойки для согласования пространств

Для согласования пространств ViT и LLM вводится преобразование $f : R^d \to R^k$, которое позволяет проектировать эмбеддинги ViT в пространство токенов LLM. Это достигается с помощью обучаемой прослойки, такой как линейная проекция, адаптер или нейронная сеть.

Обозначим преобразование:

$$T_{ViT} = \{f(z_1), f(z_2), \ldots, f(z_N)\},$$

где $f(z_i) \in \mathcal{V}_{\mathcal{LLM}}$.

На уровне абстракции результат работы прослойки f – это набор токенов T_{ViT}, которые находятся в пространстве $\mathcal{V}_{\mathcal{LLM}}$. Эти токены могут быть поданы на вход LLM:

$$\boldsymbol{T}_{LLM} = \boldsymbol{T}_{ViT} \bigcup \boldsymbol{T}_{Text},$$

где $\boldsymbol{T}_{Text}$ – текстовые токены, входящие в задачу, а $\boldsymbol{T}_{ViT}$ – объединённое представление, совместимое с LLM.

Архитектура экспериментального ViT

Для решения проблемы согласования пространств была выбрана обучаемая прослойка в виде линейного слоя. Данная прослойка преобразует выходные эмбеддинги, полученные из Vision Transformer (ViT), в пространство скрытых представлений, совместимое с крупномасштабной языковой моделью (LLM).

В экспериментальной сборке использовалась модель BLIP-2[34] в качестве ViT для обработки изображений и GPT-2 в качестве LLM для генерации текстовых токенов. Основным элементом согласования выступает линейный слой, преобразующий выходные эмбеддинги BLIP-2 в формат, подходящий для обработки GPT-2. BLIP-2 и GPT-2 предобучены. В настоящей работе производится обучение проекционной прослойки вместе с дообучением GPT-2 на наборе данных DriveLM[35] в варианте, основанном на nuScenes[36].

Архитектура

- Модуль ViT (BLIP-2): Генерирует эмбеддинги из изображений, представляющие визуальные данные в высокоразмерном пространстве.
- Линейная прослойка (aligner): Линейный слой уменьшает размерность эмбеддингов из BLIP-2 до размерности скрытого пространства GPT-2 (*hidden_size*), согласуя визуальные представления с текстовыми токенами.
- LLM (GPT-2): Языковая модель обрабатывает объединённые токены (визуальные и текстовые) для генерации текста, основанного на визуальном и текстовом входе.

Принцип работы

1. Эмбеддинги изображения, полученные из BLIP-2, преобразуются через линейный слой в скрытое пространство GPT-2.
2. Токены изображения объединяются с текстовыми токенами вопроса, формируя общий вход для GPT-2.
3. GPT-2 генерирует выходной текст, обучаясь на данных, где визуальная информация и текстовые запросы совместно определяют результат.

Набор данных DriveLM

Набор предназначен для решения задачи восприятия, прогнозирования, планирования, поведения и движения с помощью логики рассуждений, написанной человеком. Представлена задача GVQA по соединению пар QA в структуру в виде графа.

DriveLM состоит из двух отдельных наборов: DriveLM-nuScenes и DriveLM–CARLA, которые собраны на основе набора данных nuScenes и симулятора CARLA соответственно.

Набор nuScenes представляет из себя изображения, полученные с камеры на крыше автомобиля в городских условиях. Каждая сцена снята на 6 камер, направленных в разные стороны: вперед, вперед-влево, вперед-вправо, назад, назад-влево, назад-вправо.

Аннотация каждой сцены включает в себя следующие компоненты:

- Список объектов, находящихся в сцене. Структура объекта:
  - ИД объекта (c1, c2, c3, …)
  - Камера, на которой объект виден (изображения с камер не пересекаются – т. е. не имеют одинаковых объектов, для каждого объекта только одна камера)
  - Тип объекта (Транспорт, Объект ПДД)
  - Статус объекта (только для транспорта – Двигается/Стационарен)
  - Словесное описание (“Синий седан” – Транспорт, “Зеленый сигнал светофора” – Объект ПДД)
  - 2D-координаты объекта на изображении с камеры
- Словесное описание сцены, относительно автомобиля с камерами. Пример: *Эго(Автономный) -автомобиль проезжает перекресток, продолжая движение по текущей дороге..*
- Имена файлов 6-и изображений
- Список пар вопросов-ответов. Разделены на 4 группы:
  - **Поведение** (Behavior):
    Q: *Предсказать поведение эго-автомобиля.*
    A: *Эго-автомобиль едет прямо. Эго-машина едет медленно.*
  - **Восприятие** (Perception):
    Q: *Определите все элементы трафика на переднем плане, классифицируйте их, определите их статус и спрогнозируйте ограничивающую рамку вокруг каждой из них. Выход должен быть списком, отформатированным как (c, s, x1, y1, x2, y2), где c представляет категорию, s обозначает статус, а x1, y1, x2, y2 – смещения верхнего левого и нижнего правого углов рамки относительно центральной точки.*
    A: *На переднем плане много элементов дорожного движения. Информация об этих элементах дорожного движения: [(светофор, зеленый, 13.19, 382.76, 39.21, 427.96), (светофор, неизвестный, 255.86, 274.14, 279.25, 324.67), (светофор, неизвестный, 388.89, 266.58, 413.72, 318.86), (светофор, зеленый, 713.29, 306.48, 731.57, 345.14), (светофор, зеленый, 826.91, 313.56, 844.49, 351.52), (светофор, неизвестный, 399.02, 400.97, 412.73, 429.78)].*
  - **Предсказывание** (Prediction):
    Q: *Будет ли <c1, CAM_BACK,1088.3,497.5> в направлении движения эго-автомобиля?*
    A: *Нет.*
  - **Планирование** (Planning):
    Q: *Какие действия может предпринять эго-автомобиль на основе <c1, CAM_BACK,1088.3,497.5>? Зачем предпринимать это действие и какова вероятность?*

***A:** Действие заключается в том, чтобы продолжать движение с той же скоростью. Причина в том, чтобы соблюдать правила дорожного движения, что имеет большую вероятность.*

Вопросы и ответы из групп предсказывания и планирования всегда включают в себя ссылки на объекты из списка объектов, вида *<c1, CAM_BACK,1088.3,497.5>*. В остальных группах они никогда не появляются. В группе восприятия некоторые вопросы требуют позиционного определения объекта относительно 2D-координат изображения.

4. ПРОЦЕСС И РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ЭКСПЕРИМЕНТАЛЬНОЙ МОДЕЛИ VLM

В текстовую форму вопросов добавлена информация о том, с какой камеры получено текущее изображение. Вопрос также приведен к формату для обучения задаче Q&A. Пример форматированной пары вопрос-ответ: ***Q:** "Камера направлена вперед-влево. Q: Что за объекты находятся спереди справа от эго-автомобиля? A:" **A:** "Спереди справа от эго-автомобиля много заграждений и одна строительная машина."*. Текстовые данные токенизируются с использованием GPT-2-токенизатора.

Каждое изображение обрабатывается ViT-моделью (BLIP-2) для извлечения векторных представлений (токенов). Эти токены представляют визуальную информацию изображения и выравниваются с текстовыми токенами через обучаемую прослойку.

Были использованы только вопросы из категорий поведения и восприятия, т. к. две другие – предсказывание и планирование – содержат ссылки на объекты, с которыми экспериментальная сборка модели пока не может работать.

Обучение модели включает оптимизацию параметров как обучаемой прослойки, так и GPT-2 модели, что обеспечивает согласованное представление визуальных и текстовых данных в общем пространстве токенов.

Построенная модель объединяет GPT-2 (в качестве языковой модели) и обучаемую линейную прослойку для выравнивания векторных представлений изображений (полученных из ViT BLIP-2) с пространством токенов GPT-2. Обе части модели участвуют в процессе оптимизации.

Объединённые токены (визуальные и текстовые) передаются в модель. Токены, полученные из ViT, выравниваются с пространством GPT-2 с помощью обучаемой линейной прослойки. GPT-2 обрабатывает входные токены, включая преобразованные визуальные данные, для генерации ответа. Для оптимизации используется стандартная кросс-энтропийная функция потерь, встроенная в реализацию модели GPT-2 из пакета Huggingface Transformers. Потери вычисляются с использованием предсказанных выходов GPT-2 и реальных токенов ответа. Отрицательные значения маски используются для исключения padding-токенов из процесса оптимизации. Оптимизация осуществляется с помощью AdamW, где веса модели обновляются по направлению антиградиента с учётом регуляризации.

Анализ результатов обучения

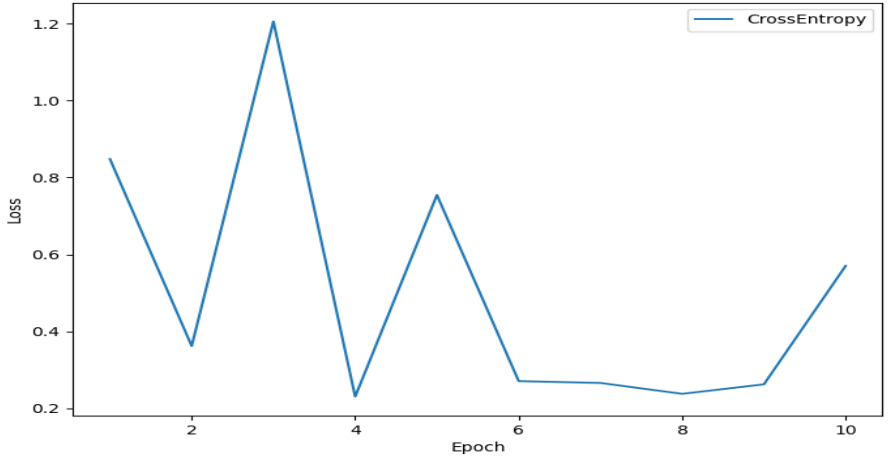

Рис. 1. График функции потерь экспериментальной модели

Обучение модели проводилось в течение 10 эпох с фиксированным набором гиперпараметров. Значение функции потерь варьировалось в диапазоне от 0.03 до 1.2 без четкой тенденции к снижению, что свидетельствует о нестабильной сходимости. На финальных этапах обучения модель достигла следующих метрик на тестовом наборе данных:

- **Перплексия (Perplexity):** 1.7096
- **Точность (Accuracy):** 86,42%

Эти показатели указывают на высокую уверенность модели в своих предсказаниях и значительное соответствие токенов в сгенерированных ответах с ожидаемыми.

5. ЗАКЛЮЧЕНИЕ

В статье представлен метод синтеза поведения когнитивного агента на основе задачи на естественном языке и объектов на изображении в комплексных сценах. Метод позволяет синтезировать план поведения агента в простых средах, но в условиях, приближенных к реальным требует доработки. Колебание значения функции потерь может указывать на нестабильность процесса оптимизации. Отсутствие снижения потерь говорит о том, что модель либо не усваивает закономерности в данных, либо уже достигла некоторого оптимума, ограниченного архитектурой модели, качеством данных или гиперпараметрами. Возможна проблема с выравниванием пространства визуальных и текстовых токенов, если обучаемая линейная прослойка не выполняет свою функцию должным образом. В будущих работах модель будет дополнена алгоритмом обучения с подкреплением, и позволит синтезировать подходящую стратегию в сложных средах.

Эксперименты показывают, что модель демонстрирует избыточную уверенность в предсказаниях и сильно склонна к повторению токенов, что будет исправлено оптимизацией самой LLM. В будущих исследованиях модель будет обучать прослойку преобразовывать патчи из ViT в токены, соответствующие словам и подсловам LLM, не изменяя ее поведение. Основное преимущество такого подхода – значительное снижение обучаемых параметров, что может помочь в достижении стабильной сходимости. Также будет протестирован подход с созданием «универсальных токенов», которые могут совместно использоваться ViT и LLM. Основной механизм заключается в разбиении выходных эмбеддингов ViT на кванты, каждый из которых сопоставляется уникальному «визуальному» токену, добавляемому в вокабуляр LLM. Это позволит расширить языковую модель для работы с изображениями.

Литература

1. *Bechon, P., Barbier, M., Grand, C., Lacroix, S., Lesire, C., & Pralet, C.* (2018). Integrating planning and execution for a team of heterogeneous robots with time and communication constraints. 1091–1097.
2. *Benjamin, D.P., Li, T., Shen, P., Yue, H., Zhao, Z., & Lyons, D.* (2018). Spatial understanding as a common basis for human-robot collaboration. Advances in Intelligent Systems and Computing. https://doi.org/10.1007/978-3-319-60384-1_3J. R. Anderson, D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, Y. Qin. An integrated theory of the mind. Psychological Review, 111(4):1036–1060, 2004.
3. *Chu, Z., Wang, Y., Zhu, F., Yu, L., Li, L., & Gu, J.* (2024). Professional Agents – Evolving Large Language Models into Autonomous Experts with Human-Level Competencies. ArXiv, abs/2402.03628.
4. *Davis, D.N., & Ramulu, S.K.* (2017). Reasoning with BDI robots: From simulation to physical environment – Implementations and limitations. Paladyn, 8(1), 39–57. https://doi.org/10.1515/pjbr-2017-0003
5. *Madl, T., Franklin, S., Chen, K., & Trappl, R.* (2018). A computational cognitive framework of spatial memory in brains and robots. Cognitive Systems Research, 47, 147–172. https://doi.org/10.1016/j.cogsys.2017.08.002
6. *Sumers, T.R., Yao, S., Narasimhan, K., & Griffiths, T.L.* (2023). Cognitive Architectures for Language Agents. Trans. Mach. Learn. Res., 2024.
7. *Emelyanov S.* and etc. (2015) Multilayer cognitive architecture for UAV control. Cognitive System Research, 34.
8. *Киселев, Г.А.* (2020). Интеллектуальная система планирования поведения коалиции робототехнических агентов с STRL архитектурой. Информационные Технологии и Вычислительные Системы. – С. 21–37. https://doi.org/10.14357/20718632200203
9. *Kiselev G., Panov A.* (2019) Hierarchical Psychologically Inspired Planning for Human-Robot Interaction Tasks. In: Ronzhin A., Rigoll G., Meshcheryakov R. (eds) Interactive Collaborative Robotics. ICR 2019. Lecture Notes in Computer Science, vol 11659. Springer, Cham.
10. *Osipov G.S., Panov A.I.* Relationships and Operations in a Sign-Based World Model of the Actor // Scientific and Technical Information Processing. 2018. No. 5.
11. *Osipov, G.S.* Sign-based representation and word model of actor. In: Yager, R., Sgurev, V., Hadjiski, M., and Jotsov, V. (eds.) 2016 IEEE 8th International Conference on Intelligent Systems (IS). pp. 2226. IEEE (2016).
12. *Leontiev A.N.* Activity Consciousness. Personality. M .: Politizdat, 1975.

13. *Bruner J.* Psychology of knowledge. Outside of direct information. M.: Progress, 1977.413 s.
14. *Kiselev G., Panov A*. Q-Learning of Spatial Actions for Hierarchical Planner of Cognitive Agents. In: Ronzhin A., Rigoll G., Meshcheryakov R. (eds) Interactive Collaborative Robotics. ICR 2020. Lecture Notes in Computer Science, (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Cham 2020, pp. 160–169. https://doi.org/10.1007/978-3-030-60337-3_16
15. *Sima, Chonghao and Renz, Katrin and Chitta, Kashyap and Chen, Li and Zhang, Hanxue and Xie, Chengen and Luo, Ping and Geiger, Andreas and Li, Hongyang.* (2024) DriveLM: Driving with Graph Visual Question Answering In: arXiv preprint arXiv:2312.14150
16. *Jintao, Xue & Zhang, Dongkun & Xiong, Rong & Wang, Yue & Liu, Eryun.* (2023). A Two-Stage Based Social Preference Recognition in Multi-Agent Autonomous Driving System. 5507–5513. 10.1109/IROS55552.2023.10341803.
17. *Tian, Ran & Li, Boyi & Weng, Xinshuo & Chen, Yuxiao & Schmerling, Edward & Wang, Yue & Ivanovic, Boris & Pavone, Marco.* (2024). Tokenize the World into Object-level Knowledge to Address Long-tail Events in Autonomous Driving. 10.48550/arXiv.2407.00959.
18. *Weng, Xinshuo & Ivanovic, Boris & Wang, Yan & Wang, Yue & Pavone, Marco.* (2024). PARA-Drive: Parallelized Architecture for Real-Time Autonomous Driving. 15449–15458. 10.1109/CVPR52733.2024.01463.
19. *Bain, A.* (1855). The senses and the intellect. London: Parker.
20. *Hull, C.L.* (1943). Principles of Behavior: An Introduction to Behavior Theory. New York: D. Appleton-Century Company.
21. *Skinner, B.F.* (1931). The concept of the reflex in the description of behavior, Ph.D. thesis, Harvard University.
22. https://github.com/neuroidss/text-generation-neurofeedback-webui
23. *Dong, Na & Zhang, Wen-qi & Gao, Zhong-ke.* (2019). Research on fuzzy PID Shared control method of small brain-controlled uav. 10.48550/arXiv.1905.12240.
24. *Осипов Г.С., Чудова Н.В., Панов А.И.* Знаковая картина мира субъекта поведения.
25. *Antol, Stanislaw & Agrawal, Aishwarya & Lu, Jiasen & Mitchell, Margaret & Batra, Dhruv & Zitnick, C. & Parikh, Devi.* (2015). VQA: Visual Question Answering. 2. 10.1109/ICCV.2015.279.
26. *Xie, S., Sun, C., Huang, J., Tu, Z., and Murphy, K.* Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European conference on computer vision (ECCV), pp. 305–321, 2018.
27. *Gupta, A., Pacchiano, A., Zhai, Y., Kakade, S.M., and Levine, S.* Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity, 2022.
28. *Chu, K., Zhao, X., Weber, C., Li, M., and Wermter, S.* Accelerating reinforcement learning of robotic manipulations via feedback from large language models. arXiv preprint arXiv: 2311.02379, 2023.
29. *Radford A. et al.* Language models are unsupervised multitask learners // OpenAI blog. – 2019. – Т. 1. – № . 8. – P. 9.
30. T*ouvron H.* et al. Llama: Open and efficient foundation language models // arXiv preprint arXiv:2302.13971. – 2023.
31. *Tan H., Bansal M.* Lxmert: Learning cross-modality encoder representations from transformers // arXiv preprint arXiv:1908.07490. – 2019.
32. *Li X.* et al. Oscar: Object-semantics aligned pre-training for vision-language tasks // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. – Springer International Publishing, 2020. – С. 121–137.
33. *Zhang P.* et al. Vinvl: Revisiting visual representations in vision-language models // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2021. – С. 5579–5588.

34. *Li J.* et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models // International conference on machine learning. – PMLR, 2023. – C. 19730–19742.
35. *Sima C.* et al. Drivelm: Driving with graph visual question answering // arXiv preprint arXiv:2312.14150. – 2023.
36. *Caesar H.* et al. nuscenes: A multimodal dataset for autonomous driving // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2020. – C. 11621–11631.

Method for Creating Behavior of Cognitive Agents Based on Multimodal Signal Processing

Daniil A. Weizenfeld*
Peoples' Friendship University of Russia (RUDN), Moscow, Russia
ORCID: https://orcid.org/0000-0002-2787-0714
e-mail: veicenfeld@isa.ru

Gleb A. Kiselev**
Federal Research Center "Informatics and Management"
Russian Academy of Sciences (FRC CSC RAS)
Peoples' Friendship University of Russia (RUDN), Moscow, Russia
ORCID: https://orcid.org/0000-0001-9231-8662
e-mail: kiselev@isa.ru

The paper considers the problem of predicting the agent's activity based on the text description of the task and visual analysis of the environment. An update of the approaches of classical cognitive architecture is proposed, allowing its application in a real environment. An addition to the semiotic method of symbolic designation with the author's neural network mechanism for linking vectors of text and visual spaces is developed. A number of experiments with the obtained model in a complex environment of a car driving emulator are conducted.

Keywords: activity-related experience, quality of motivation, self-determination theory, intrinsic motivation, extrinsic motivation, academic motivation.

Acknowledgements. This publication has been supported by the RUDN University Strategic Academic Leadership Program, project No. 021934-0-000.
The research was carried out using the infrastructure of the Shared Research Facilities «High Performance Computing and Big Data» (CKP «Informatics») of FRC CSC RAS (Moscow).

For citation:
Weizenfeld D., Kiselev G. Method for Creating Behavior of Cognitive Agents Based on Multimodal Signal Processing. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2024. Vol. 14, no. 4, pp. 45–62. DOI: https://doi.org/10.17759/mda.2024140403 (In Russ., abstr. in Engl.).

***Daniil A. Weizenfeld,** Master's Degree Student, Peoples' Friendship University of Russia (RUDN), Moscow, Russia, ORCID: https://orcid.org/0000-0002-2787-0714, e-mail: veicenfeld@isa.ru

****Gleb A. Kiselev,** Candidate of Technical Sciences, Researcher, Federal Research Center "Informatics and Management", Russian Academy of Sciences (FRC CSC RAS); Senior Lecturer, Peoples' Friendship University of Russia (RUDN), Moscow, Russia, ORCID: https://orcid.org/0000-0001-9231-8662, e-mail: kiselev@isa.ru

References

1. Bechon, P., Barbier, M., Grand, C., Lacroix, S., Lesire, C., & Pralet, C. (2018). Integrating planning and execution for a team of heterogeneous robots with time and communication constraints. pp. 1091–1097.
2. Benjamin, D. P., Li, T., Shen, P., Yue, H., Zhao, Z., & Lyons, D. (2018). Spatial understanding as a common basis for human-robot collaboration. Advances in Intelligent Systems and Computing. https://doi.org/10.1007/978-3-319-60384-1_3J. R. Anderson, D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, Y. Qin. An integrated theory of the mind. Psychological Review, 111(4):1036–1060, 2004.
3. Chu, Z., Wang, Y., Zhu, F., Yu, L., Li, L., & Gu, J. (2024). Professional Agents – Evolving Large Language Models into Autonomous Experts with Human-Level Competencies. ArXiv, abs/2402.03628.
4. Davis, D. N., & Ramulu, S. K. (2017). Reasoning with BDI robots: From simulation to physical environment – Implementations and limitations. Paladyn, 8(1), 39–57. https://doi.org/10.1515/pjbr-2017-0003
5. Madl, T., Franklin, S., Chen, K., & Trappl, R. (2018). A computational cognitive framework of spatial memory in brains and robots. Cognitive Systems Research, 47, 147–172. https://doi.org/10.1016/j.cogsys.2017.08.002
6. Sumers, T.R., Yao, S., Narasimhan, K., & Griffiths, T.L. (2023). Cognitive Architectures for Language Agents. Trans. Mach. Learn. Res., 2024.
7. Emelyanov S. and etc. (2015) Multilayer cognitive architecture for UAV control. Cognitive System Research, 34.
8. Kiselev, G.A. (2020). Intellektual'naya sistema planirovaniya povedeniya koalitsii robototekhnicheskikh agentov s STRL arkhitekturoi. Informatsionnye Tekhnologii i Vychislitel'nye Sistemy. – pp. 21–37. https://doi.org/10.14357/20718632200203
9. Kiselev G., Panov A. (2019) Hierarchical Psychologically Inspired Planning for Human-Robot Interaction Tasks. In: Ronzhin A., Rigoll G., Meshcheryakov R. (eds) Interactive Collaborative Robotics. ICR 2019. Lecture Notes in Computer Science, vol 11659. Springer, Cham.
10. Osipov G.S., Panov A.I. Relationships and Operations in a Sign-Based World Model of the Actor // Scientific and Technical Information Processing. 2018. No. 5.
11. Osipov, G.S. Sign-based representation and word model of actor. In: Yager, R., Sgurev, V., Hadjiski, M., and Jotsov, V. (eds.) 2016 IEEE 8th International Conference on Intelligent Systems (IS). pp. 2226. IEEE (2016).
12. Leontiev A.N. Activity Consciousness. Personality. M.: Politizdat, 1975.
13. Bruner J. Psychology of knowledge. Outside of direct information. M.: Progress, 1977.413 s.
14. Kiselev G., Panov A. Q-Learning of Spatial Actions for Hierarchical Planner of Cognitive Agents. In: Ronzhin A., Rigoll G., Meshcheryakov R. (eds) Interactive Collaborative Robotics. ICR 2020. Lecture Notes in Computer Science, (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, Cham 2020, pp. 160–169. https://doi.org/10.1007/978-3-030-60337-3_16
15. Sima, Chonghao and Renz, Katrin and Chitta, Kashyap and Chen, Li and Zhang, Hanxue and Xie, Chengen and Luo, Ping and Geiger, Andreas and Li, Hongyang. (2024) DriveLM: Driving with Graph Visual Question Answering In: arXiv preprint arXiv:2312.14150
16. Jintao, Xue & Zhang, Dongkun & Xiong, Rong & Wang, Yue & Liu, Eryun. (2023). A Two-Stage Based Social Preference Recognition in Multi-Agent Autonomous Driving System. 5507–5513. 10.1109/IROS55552.2023.10341803.

17. Tian, Ran & Li, Boyi & Weng, Xinshuo & Chen, Yuxiao & Schmerling, Edward & Wang, Yue & Ivanovic, Boris & Pavone, Marco. (2024). Tokenize the World into Object-level Knowledge to Address Long-tail Events in Autonomous Driving. 10.48550/arXiv.2407.00959.
18. Weng, Xinshuo & Ivanovic, Boris & Wang, Yan & Wang, Yue & Pavone, Marco. (2024). PARA-Drive: Parallelized Architecture for Real-Time Autonomous Driving. 15449–15458. 10.1109/CVPR52733.2024.01463.
19. Bain, A. (1855). The senses and the intellect. London: Parker.
20. Hull, C. L. (1943). Principles of Behavior: An Introduction to Behavior Theory. New York: D. Appleton-Century Company.
21. Skinner, B. F. (1931). The concept of the reflex in the description of behavior, Ph.D. thesis, Harvard University.
22. https://github.com/neuroidss/text-generation-neurofeedback-webui
23. Dong, Na & Zhang, Wen-qi & Gao, Zhong-ke. (2019). Research on fuzzy PID Shared control method of small brain-controlled uav. 10.48550/arXiv.1905.12240.
24. Osipov G.S., Chudova N.V., Panov A.I. Znakovaya kartina mira sub”ekta povedeniya.
25. Antol, Stanislaw & Agrawal, Aishwarya & Lu, Jiasen & Mitchell, Margaret & Batra, Dhruv & Zitnick, C. & Parikh, Devi. (2015). VQA: Visual Question Answering. 2. 10.1109/ICCV.2015.279.
26. Xie, S., Sun, C., Huang, J., Tu, Z., and Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European conference on computer vision (ECCV), pp. 305–321, 2018.
27. Gupta, A., Pacchiano, A., Zhai, Y., Kakade, S. M., and Levine, S. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity, 2022.
28. Chu, K., Zhao, X., Weber, C., Li, M., and Wermter, S. Accelerating reinforcement learning of robotic manipulations via feedback from large language models. arXiv preprint arXiv: 2311.02379, 2023.
29. Radford A. et al. Language models are unsupervised multitask learners // OpenAI blog. – 2019. – T. 1. – № . 8. – P. 9.
30. Touvron H. et al. Llama: Open and efficient foundation language models // arXiv preprint arXiv:2302.13971. – 2023.
31. Tan H., Bansal M. Lxmert: Learning cross-modality encoder representations from transformers // arXiv preprint arXiv:1908.07490. – 2019.
32. Li X. et al. Oscar: Object-semantics aligned pre-training for vision-language tasks // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. – Springer International Publishing, 2020. – P. 121–137.
33. Zhang P. et al. Vinvl: Revisiting visual representations in vision-language models // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2021. – P. 5579–5588.
34. Li J. et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models // International conference on machine learning. – PMLR, 2023. – P. 19730–19742.
35. Sima C. et al. Drivelm: Driving with graph visual question answering // arXiv preprint arXiv:2312.14150. – 2023.
36. Caesar H. et al. nuscenes: A multimodal dataset for autonomous driving // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2020. – P. 11621–11631.

Получена 15.11.2024 Received 15.11.2024
Принята в печать 02.12.2024 Accepted 02.12.2024