Инженерный вестник Дона, №2 (2026)
ivdon.ru/ru/magazine/archive/n2y2026/10785

Мультиагентный контроллер покрытия: метод глубокого обучения с подкреплением для решения задачи планирования пути покрытия в мультиагентных системах

С.Ю. Луканов

Псковский государственный университет

Аннотация: В работе представлен мультиагентный контроллер покрытия (Multi-Agent Coverage Controller – MACC) – специализированный метод глубокого обучения с подкреплением для решения задачи планирования пути покрытия в мультиагентных системах. Метод решает ключевые проблемы задачи планирования пути покрытия: редкие и шумные вознаграждения, высокую дисперсию градиента, сложность распределения заслуг между агентами и необходимость масштабирования на переменное число агентов. Метод интегрирует определенный набор механизмов: адаптивная ширина интервала обрезки, шлюзование модулированного преимущества, контрфактический базис централизованного критика и механизм многоголового самовнимания с маской присутствия. Представлены теоретические свойства метода, подтверждающие стабильность оптимизации и снижение дисперсии градиентных оценок. Проведен комплексный абляционный анализ, демонстрирующий вклад каждого механизма в координацию агентов, пространственное распределение траекторий и итоговую скорость покрытия. Эксперименты на наборе спутниковых карт показывают, что MACC достигает значительного увеличения полноты и скорости покрытия по сравнению с базовой конфигурацией, обеспечивая наилучшие результаты при одновременном использовании всех интегрированных механизмов.
Ключевые слова: мультиагентная система, планирование пути покрытия, глубокое обучение с подкреплением, адаптивная ширина интервала обрезки, шлюзование модулированного преимущества, контрфактический базис, механизм многоголового самовнимания, координация агентов.

Введение

Задача планирования пути покрытия в мультиагентной среде представляет собой тип коллективного поведения, при котором группа агентов должна скоординированно покрыть заданную территорию или пространство. Интерес к подобным задачам растет, поскольку мультиагентные системы активно внедряются в логистике, мониторинге, сельском хозяйстве, складской автоматизации и робототехнических комплексах.

Решение подобных задач с помощью глубокого обучения с подкреплением сталкивается с рядом серьезных вызовов. Во-первых,

мультиагентная среда является нестационарной с точки зрения каждого отдельного агента: поведение окружения изменяется по мере обновления стратегий других агентов. Это затрудняет обучение, так как стандартные подходы предполагают стационарность среды. Во-вторых, присутствует проблема распределения вклада агентов – глобальная награда обычно выдается для всей команды, и отдельным агентам трудно оценить свой вклад в общий результат. В-третьих, число агентов может быть переменным, а совместное пространство состояний и действий растет экспоненциально с увеличением числа агентов, что усложняет масштабирование алгоритма.

Актуальность развития специализированных методов подтверждается широким интересом к подобным задачам в современных исследованиях. В робототехнике отмечается важность обобщающих стратегий и снижения числа онлайн-взаимодействий, что привело к разработке гибридных подходов, объединяющих генеративные модели и глубокое обучение с подкреплением – такие как двухфазный метод обучения роботов по демонстрациям, использующий диффузионную модель и алгоритм оптимизации проксимальной политики (Proximal Policy Optimization – PPO) [1]. В прикладных областях, где требуется пространственная оптимизация и принятие решений, методы глубокого обучения с подкреплением демонстрируют свою эффективность: так, в задаче оптимизации дислокации мест рубок в лесозаготовительной отрасли была показана возможность использования PPO для автоматизации выбора оптимальных конфигураций лесосек на больших территориях, что подтверждает востребованность подобных алгоритмов в реальных производственных процессах [2]. Кроме того, исследования в области оптимизации производственных графиков показывают, что гибридизация глубокого обучения с подкреплением с методами поиска обеспечивает существенное улучшение качества решений в задачах планирования с большим пространством состояний, что

подчеркивает эффективность подходов, разрабатывающих адаптивные стратегии поведения агентов в сложных средах [3].

Популярные на текущий момент методы (такие как независимые агенты, централизованное обучение с децентрализованным выполнением, методы на основе декомпозиции и факторизации функции ценности) разрабатывались как универсальные решения для широкого круга задач. Однако в задаче покрытия, где вознаграждения редки, шумны и зависят от совместных действий группы агентов, их применение оказывается ограниченным. Это требует разработки методов, специально адаптированных к особенностям задачи планирования пути покрытия.

Для преодоления указанных проблем настоящая работа предлагает метод глубокого обучения с подкреплением: мультиагентный контроллер покрытия.

В работе последовательно изложены: постановка задачи, обоснование выбора архитектурных и практических решений с их подробным описанием, теоретические свойства разработанного метода, результаты экспериментального тестирования разработанного метода с абляционным анализом.

Постановка задачи

Задача планирования пути покрытия может быть формализована следующим образом. Пусть $A \subset \mathbb{R}^2$– вся область. $O \subset A$ – множество препятствий, которые необходимо избегать. $U \subset A$ – целевая область, которую необходимо покрыть. $F \subset A \setminus O$ – доступная для движения область. $N \in \mathbb{N}$– число агентов. Для каждого агента $i \in \{1, \ldots, N\}$ задается:

- начальная позиция $x^0_i \in F$;
- функции движения $x_i(t) : [0, T] \to F$, где $x_i(t)$ – положение агента i в момент времени t;

- зона охвата $S_i(x) \subset F$ – подмножество области, покрываемое агентом i, находящимся в позиции $x \subset F$.

Совокупная зона покрытия C всеми агентами в момент времени t определяется формулой:

$$C(t) = \bigcup_{i=1}^{N} S_i\left(x_i(t)\right), \quad (1)$$

где S – зона охвата, x – позиция агента, t – временной шаг, i – индекс агента, N – количество агентов.

Совокупная покрытая область за все время E определяется формулой:

$$E \subseteq \bigcup_{t\in[0,T]} C(t) \Leftrightarrow U \setminus O \subseteq \bigcup_{t\in[0,T]} \bigcup_{i=1}^{N} S_i\left(x_i(t)\right), \quad (2)$$

где C – совокупная зона покрытия, S – зона охвата, x – позиция агента, t – временной шаг, i – индекс агента, N – количество агентов, U – целевая область, O – множество препятствий, T – последний временной шаг траектории агента.

Цель: найти все траектории $x_1(t), \ldots, x_N(t)$ такие, что:

- обеспечивается полное покрытие области: $U \subseteq E$;
- происходит избегание препятствий: $x_i(t) \in F$, $\forall t \in [0, T]$, $\forall i \in \{1, \ldots, N\}$;
- происходит оптимизация времени: $\min\limits_{x_1(t),\ldots,x_N(t)} T$.

Данная формализация обобщает дискретные и непрерывные методы. При необходимости могут быть введены дополнительные ограничения, такие как скорость или угол поворота, и эквивалентные критерии.

Представим задачу мультиагентного планирования пути покрытия как частично наблюдаемую марковскую среду M в следующем виде:

$$M = \left(S, \{A_i\}_{i=1}^{N}, P, \{r_i\}_{i=1}^{N}, \{O_i\}_{i=1}^{N}, \gamma\right), \quad (3)$$

где S – пространство глобальных состояний, A – пространство действий, P – динамика среды, r – мгновенное вознаграждение, O – локальное наблюдение,

$\gamma \in (0, 1)$ – коэффициент дисконтирования, i – индекс агента, N – количество агентов.

Целевым критерием задачи является минимизация времени достижения требуемого уровня покрытия. Формально задачу можно записать как минимизацию ожидаемого времени завершения эпизода:

$$\min_{\theta} \mathbb{E}_{\tau \sim \pi_\theta} [\tau(\pi_\theta)], \quad (4)$$

где τ – время завершения эпизода, π – стратегия агента, θ – параметры стратегии, $\mathbb{E}[X]$ – математическое ожидание случайной величины X.

Для совместимости с on-policy методами, удобно перейти к эквивалентной формулировке через вознаграждение: установим награду за шаг $r_t = -1$, пока покрытие не достигнуто, и $r_t = 0$ после успешного завершения. Тогда минимизация ожидаемого времени эквивалентна максимизации дисконтированной суммы отрицательных вознаграждений:

$$\max_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [\sum_{t=0}^{\tau-1} \gamma^t (-1)], \quad (5)$$

где J – целевая функция, τ – время завершения эпизода, π – стратегия агента, θ – параметры стратегии, t – временной шаг, $\gamma \in (0, 1)$ – коэффициент дисконтирования, $\mathbb{E}[X]$ – математическое ожидание случайной величины X.

Поскольку эпизод гарантированно завершается при достижении полного покрытия, суммирование наград является конечным.

Для проведения корректного теоретического анализа работы алгоритма и получения формально обоснованных свойств градиентного обновления вводится набор стандартных допущений, широко применяемых в исследованиях методов градиента стратегии и подходов актор-критик. Эти допущения обеспечивают существование и ограниченность необходимых величин, корректность градиентных оценок и стабильность процедуры обучения. Ниже перечислены используемые предпосылки:

• вознаграждения ограничены: $\forall s, a, i, |r_i(s,a)| \leq R_{max}$, где R_{max} – максимально возможное значение вознаграждения;

• стратегии параметризованы гладко по параметрам θ, для логарифма стратегии $log\pi_\theta(a|o)$ существуют и ограничены градиенты по θ;

• критик имеет ограниченную аппроксимационную ошибку и оптимизируется так, что среднеквадратическая ошибка остается ограниченной малым ε;

• метод работает в онлайн режиме и пакеты данных собираются по старой стратегии π^{old} обновления происходят относительно этих пакетов.

Метод мультиагентного контроллера покрытия

Метод мультиагентного контроллера покрытия (Multi-Agent Coverage Controller – MACC) представляет собой модифицированный подход, основанный на PPO [4] и адаптированный к специфике задачи мультиагентного покрытия: редкими вознаграждениями, сильной зависимостью действий агентов, необходимостью устойчивой оптимизации времени эпизода и переменным числом участников.

Метод интегрирует ряд ключевых механизмов, направленных на снижение дисперсии градиента и повышение стабильности обучения:

• адаптивная ширина интервала обрезки (Adaptive Clipping – AC) – контролирует величину обновления стратегии ε, сопоставимо с идеей регионов доверия. Использование данного механизма мотивировано следующим: при фиксированном значении ε возникают либо слишком ограниченные обновления, приводящие к медленной сходимости, либо частые большие скачки, приводящие к нестабильности обучения. Адаптация по эмпирической KL-дивергенции дает регулировку шага, управляемую данными;

• шлюзование модулированного преимущества (Advantage Modulation Gating – AMG) – подавление выбросов в оценках функции преимущества. В

постановке задачи через минимизацию времени значения преимуществ часто бывают отрицательными. Модификация преимущества уменьшает вклад экстремальных по модулю значений (как больших отрицательных, так и редких положительных наград), снижая дисперсию;

- контрфактический базис в централизованном критике (Counterfactual Baseline – CB) – улучшает распределение заслуг агентов. CB учитывает влияние действий других агентов и вычисляет контрфактическую поправку, уменьшая дисперсию оценок градиента и не вводя смещения при корректной реализации;
- механизм многоголового самовнимания с маской присутствия (Multi-Head Self-Attention – MHSA) – поддерживает переменное количество агентов.

В сочетании перечисленные компоненты направлены на улучшение сходимости, стабильности и эффективности конечных стратегий в задачах минимизации времени покрытия. Включение каждого механизма мотивировано решением конкретных проблем задачи покрытия:

- редкая и шумная награда приводит к нестабильным оценкам преимущества и завышенным градиентам;
- сильная координационная зависимость приводит к высокой дисперсии функции ценности;
- переменное число агентов приводит к необходимости в способности корректно работать с любым количеством агентов и масштабируемости;
- онлайн-режим обучения приводит к повышенной чувствительности к крупным обновлениям стратегий.

Приведем краткие описания обоснований каждого компонента и его связь со спецификой задачи.

Адаптивная ширина интервала обрезки: обычный PPO использует фиксированное значение ε, что создает компромисс между стабильностью и скоростью обучения [4]. В задачах покрытия, где награды редкие, слишком

малый ε приводит к медленной адаптации, а слишком большой – к резкому росту KL-дивергенции и последующей деградации качества стратегии. В MACC параметр ε регулируется через обратную связь по эмпирической KL-дивергенции, аналогично идее доверительных регионов в алгоритме оптимизации политики доверительного региона (Trust Region Policy Optimization – TRPO) [5]. Это обеспечивает: стабилизацию обновления стратегии; предотвращение больших «скачков», критичных при минимизации времени покрытия; повышение устойчивости оптимизации в онлайн режиме при шумных градиентах.

Шлюзование модулированного преимущества: в задачах покрытия значения преимущества часто имеют тяжелые хвосты – длительные последовательности отрицательных вознаграждений и редкие крупные положительные сигналы. Это ведет к высокой дисперсии стохастического градиента. Гладкая модуляция преимуществ уменьшает вклад экстремумов; масштабная поправка возвращает среднюю величину, что снижает риск систематического смещения. Это особенно важно при оценке времени покрытия, где редкие события сильно влияют на значение функции преимущества.

Контрфактический базис: использование СВ восходит к алгоритму контрфактический мультиагент (Counterfactual Multi-Agent – COMA) [6] и доказано, что минимизирует условную дисперсию градиента в классе базисных функций, зависящих от (s, a_{-i}), где s – глобальное состояние среды, a_{-i} – совокупность действий всех агентов, кроме агента с индексом i. Для задач покрытия это особенно критично, так как вклад агента в общее время эпизода зависит от его взаимодействий с локальными соседями. Контрфактический базис сохраняет несмещенность градиента и, при корректном аппроксиматоре функции, дает значительное снижение дисперсии оценок. Таким образом, контрфактический базис обеспечивает

несмещенность оценки градиента, уменьшает дисперсию Q-функции, существенно улучшает кредитное присвоение (credit assignment).

Механизм многоголового самовнимания с маской присутствия: подходы на основе самовнимания зарекомендовали себя как наиболее эффективные в масштабировании на переменное число агентов [7, 8]. Включение маски присутствия и отслеживание числа активных агентов *n_active* обеспечивает инвариантность сигналов и работы критика относительно перестановок агентов при применении суммирования или усреднения по *n_active.* Нормализованное значение количества *n_active* в составе наблюдения предоставляет стратегии контекстную информацию о пространственной плотности агентов, что может способствовать ускоренной адаптации. Однако данное включение сопровождается передачей глобальной информации, что формирует компромисс между повышением адаптивности и сохранением строго локальной природы стратегии. При необходимости обеспечения полной децентрализации следует исключить количество активных агентов из локального наблюдения, а в случае приоритета ускоренной адаптации – включить указанную величину с соответствующей нормализацией.

Теоретические свойства метода мультиагентного контроллера покрытия

Свойство 1. Адаптивная ширина интервала обрезки стабилизирует KL-дивергенцию.

Пусть в обучении на каждом шаге используется адаптивная ширина интервала обрезки (ε_t) по следующему правилу:

$$\varepsilon_t = clamp\left(\varepsilon_0 * \frac{KL\_target}{KL\_batch}, \varepsilon_{min}, \varepsilon_{max}\right), \quad (6)$$

где *clamp* – функция, ограничивающая значение в заданном диапазоне, $KL\_batch = \mathbb{E}_{(o,a)\sim data}[log\pi^{old}(a|o) - log\pi(a|o)]$ – эмпирическая оценка $KL(\pi^{old} || \pi)$ перед обновлением, *KL_target* – целевая KL-дивергенция, ε_0 –

базовое значение ε, ε_{min} – минимальное значение ε, ε_{max} – максимальное значение ε.

При стандартных допущениях гладкости и малости шага градиента, аналогичных использующимся в анализе PPO, при достаточной регулярности и при маленьких шагах градиента существует стационарный режим, в котором $\mathbb{E}[KL\_batch] \approx KL\_target$.

Пояснение. Этот механизм эквивалентен отрицательной обратной связи, используемой в TRPO [5], где шаг ограничивается доверительным регионом. Ограничение предотвращает резкие изменения ε_t и устраняет бифуркации поведения обновлений, что особенно важно для онлайн методов при редких наградах.

Свойство 2. Контрфактический базис (b) сохраняет несмещенность градиента и минимизирует дисперсию.

Пусть справедливо, что:

$$b_i(s, a_{-i}) = \mathbb{E}_{a_i' \sim \pi_i(\cdot|o_i)}\left[Q\left(s, (a_i', a_{-i})\right)\right], \tag{7}$$

где b – контрфактический базис, Q – оценка Q-функции, i – индекс агента, s – глобальное состояние среды, a_{-i} – совокупность действий всех агентов, кроме агента с индексом i, π – стратегия агента, a_i' – альтернативное действие агента, o – наблюдение агента, $\mathbb{E}[X]$ – математическое ожидание случайной величины X.

Тогда $b_i(s, a_{-i})$ не зависит от фактического действия a_i. Следовательно, для стандартной стохастической градиентной оценки стратегии верно:

$$\mathbb{E}_{a_i \sim \pi_i}[\nabla_\theta log \pi_i(a_i \mid o_i) b_i(s, a_{-i})] = 0, \tag{8}$$

где π – стратегия агента, θ – параметры стратегии, i – индекс агента, s – глобальное состояние среды, a – действие агента, o – наблюдение агента, a_{-i} – совокупность действий всех агентов, кроме агента с индексом i, b –

контрфактический базис, $\mathbb{E}[X]$ – математическое ожидание случайной величины X.

Отсюда вытекает следующее: вычитание b_i из Q не смещает ожидаемую величину градиентной оценки. Результат справедлив при условии, что аппроксиматор Q является несмещенным или асимптотически консистентным.

Пояснение. Результат обобщает классическую теорию стохастического градиента с базисами-смягчителями, применимую как в REINFORCE, так и в актор-критик методах. В мультиагентной задаче покрытия это существенно снижает вариативность оценок и ускоряет обучение.

Свойство 3. AMG снижает дисперсию при контролируемой функции Липшица.

В общем виде нелинейное преобразование преимуществ g(A) может уменьшать или увеличивать дисперсию градиента выборки. Для получения однозначного теоретического преимущества необходимо ограничить гладкость g. Пусть $g: \mathbb{R} \to \mathbb{R}$ – функция с константой Липшица L, $|g(x) - g(y)| \leq L|x - y|$ для всех x, y. Тогда для случайно величины A с конечной дисперсией справедливо, что $D(g(A)) \leq L^2 D(A)$, где $D[X]$ – дисперсия случайной величины X. В частности, если $L \leq 1$, то $D(g(A)) \leq D(A)$.

Пояснение. Это следует из свойств отображений Липшица случайных величин. Рассмотрим функцию $g_\beta(x) = x\sigma(\beta x)$, где g_β – функция сглаживания и масштабируемого подавления выбросов AMG, x – входное значение преимущества, которое подается на функцию сглаживания, σ – сигмоидная функция активации, β – коэффициент, управляющий степенью модификации преимущества. При использовании данной функции в MACC, она может иметь $L > 1$, однако корректировка масштаба на пакете данных компенсирует возможное смещение, аналогично техникам пакетной нормализации [9].

Свойство 4. При использовании CB совместно с AMG сохраняется несмещенность пакетного градиента стратегии.

Пусть: $m = \overline{s \cdot g(A^{CF})}$, коэффициент s определяется так, что среднее на пакете данных совпадает: $\mathbb{E}_{batch}[\mathrm{m}] = \mathbb{E}_{batch}[A^{CF}]$, где A^{CF} – контрфактическое преимущество, g – функция AMG, s – коэффициент масштабирования, $\mathbb{E}[X]$ – математическое ожидание случайной величины X. Тогда оценка стохастического градиента на пакете остается несмещенной.

Пояснение. Хотя несмещенность строго гарантируется только на конечной выборке, в пределе больших пакетов оценка стремится к истинному значению. Такой подход используется во многих стабилизирующих надстройках над актор-критик методами [9, 10].

Эксперименты и результаты

В данном разделе проверяется эффективность предлагаемого подхода. Сначала приводится описание тестовой среды, затем общая схема тестирования и, наконец, приведены результаты экспериментального тестирования.

Тестовая среда представляет собой набор предварительно обработанных спутниковых карт водоемов, на которых водные массивы – целевые зоны, помечены зеленым цветом. Агенты имеют радиус покрытия и получают наблюдения с двух камер: одна охватывает всю местность, вторая охватывает территорию в непосредственной близости агента. На своих камерах агенты отображаются красным цветом, другие агенты отображаются синим. Границы и препятствия отображаются белым цветом. Препятствия движутся по прямой траектории в одном направлении. За один шаг агент может переместиться на одну ячейку в восьми направлениях. Вознаграждение на шаге i определяется по формуле:

$$r_i = \frac{N_i * D}{T} - P, \quad (9)$$

где N_i – количество покрытых ячеек на шаге i, D – награда за покрытую ячейку, T – количество целевых ячеек на карте, P – временной штраф.

При попытке выхода за границы карты или при столкновении с препятствием агент получает дополнительный штраф и деактивируется. Эпизод завершается в одном из трех случаев:

- при завершении покрытия;
- при деактивации всех агентов;
- при достижении максимального количества шагов.

Значения использованных параметров конфигурации среды с пояснениями представлены в таблице №1.

Таблица №1

Параметры конфигурации обучающей среды

Параметр	Значение	Пояснение
cameraResolution	100	Разрешение камеры в пикселях для сбора визуальных наблюдений в пикселях
mapSize	300	Размер стороны карты в пикселях
wallPenalty	1,0	Штраф за столкновение с границей области
discoveryReward	5	Награда за покрытие одной ячейки
timePenalty	0,0005	Штраф за один временной шаг
visionRange	3	Радиус покрытия области агентом в пикселях
maxSteps	6000	Максимальное число шагов в эпизоде
numAgents	4	Количество агентов
numObstacles	4	Количество препятствий

Агенты обучаются на наборе из двухсот карт. Тестирование и сбор статистики темпов покрытия и координат агентов осуществляется на отдельном наборе из ста карт. Во время обучения агентов собираются стандартные метрики, такие как среднее вознаграждение, средняя длина

эпизода, ошибки аппроксиматоров, дисперсии вознаграждений. После сбора статистик вычисляется ряд метрик, включающий среднюю скорость покрытия, среднее время завершения эпизода, среднее межагентное расстояние, которое высчитывается как среднее евклидово расстояние между парами агентов.

Для оценки вклада отдельных архитектурных и функциональных компонентов предложенного метода был проведен комплексный абляционный анализ. Его цель – выявить, как каждый механизм влияет на стабильность обучения, способность агентов к координации, пространственное расхождение траекторий и итоговую скорость покрытия области. Для всех конфигураций использовались идентичные гиперпараметры, включая архитектуры актора и критика, структуру наблюдений и параметры среды, что обеспечивает корректное сравнение результатов между экспериментами.

Абляционный анализ включал последовательное включение и отключение четырех механизмов метода MACC. Все метрики вычислялись на тестовом наборе карт и усреднялись по 10 запускам с разными начальными инициализациями, что снижает влияние случайности на итоговые оценки. Результаты абляционного анализа представлены в таблице №2.

Таблица №2

Результаты абляционного анализа

Конфигурация	Средний процент покрытия (% от общего количества целевых ячеек)	Средняя скорость покрытия (ячеек/шаг)	Средняя межагентная дистанция (ячеек)	Дисперсия вознаграждения	Средняя сходимость (полных эпизодов)

Base	86%	12,9	73	2,63	316
AC	89%	13,3	80	1,55	411
AMG	88%	13,1	79	1,62	396
CB	94%	15,2	131	2,3	260
MHSA	89%	13,6	84	2,65	311
AC + AMG	91%	13,7	84	1,26	398
AC + AMG + MHSA	92%	14,2	83	1,88	382
AC + AMG + CB	95%	15,6	135	1,42	296
AC + AMG + CB + MHSA	98%	16,3	158	1,2	255

Полученные результаты позволяют вкратце обсудить роль каждого механизма в методе MACC.

AC снижает вероятность резких обновлений стратегии, ограничивая рост KL-дивергенции. Более стабильные обновления стратегии приводят к тому, что агентам становится тяжелее сойтись к одинаковой стратегии и синхронизироваться. Это в свою очередь объясняет небольшое повышение межагентной дистанции и умеренный рост покрытия: менее агрессивные стратегии ведут к более равномерному распределению маргинальных направлений движения.

AMG подавляет экстремальные значения преимущества, характерные для задач с редкими положительными наградами и длинными сериями отрицательных наград. Это снижает шум в обновлениях и приводит к тому же эффекту: агенты вероятнее придерживаются устойчивых, но различающихся стратегий, что уменьшает концентрацию агентов в одной области. Увеличение межагентной дистанции объясняется уменьшением стохастичности стратегии. Умеренное улучшение покрытия связано с сокращением числа ситуаций, когда агенты многократно перекрывают ранее посещенные участки.

СВ является наиболее значимым механизмом: он уменьшает дисперсию глобального градиента за счет корректного распределения заслуг между агентами. В задачах покрытия это особенно важно: агенты перестают дублировать роль друг друга и начинают распределять территории более эффективно. СВ дает максимальный рост межагентной дистанции и полноты покрытия.

MHSA улучшает способность критика воспринимать пространственные и поведенческие взаимосвязи агентов. Он поддерживает переменное число агентов и позволяет учитывать удаленные зависимости. Однако без СВ его эффект ограничен: MHSA указывает на то, с кем нужно координироваться, однако не сообщает как именно это делать без контрфактического распределения заслуг.

Абляционный анализ показывает, что вклад различных компонентов MACC существенно различается. Наибольшее влияние на качество покрытия и координацию оказывают СВ и MHSA, что согласуется с теоретическим анализом распределения заслуг и обработки межагентных зависимостей. AC и AMG обеспечивают важную стабилизацию стохастического градиента и уменьшают риск крупных изменений стратегии, что приводит к более плавным траекториям и уменьшению перекрытий исследуемых областей.

Стоит отдельно отметить два важных наблюдения:

- увеличение межагентной дистанции при использовании AC и AMG не связано с координационным поведением, а является побочным эффектом сглаживание траекторий поведения и уменьшение резких стохастических переключений в стратегии;
- скорость покрытия в отдельных конфигурациях может превышать базовый уровень более чем на 10–20 процентных пунктов, что объясняется уменьшением повторных визитов в уже покрытые области – ключевым фактором повышения эффективности в задачах покрытия.

Наилучший результат достигается при одновременном включении всех четырех механизмов. Такой вариант обеспечивает минимальную дисперсию градиента, ускоренную сходимость, наилучшее пространственное распределение агентов и максимальную полноту покрытия.

Заключение

В работе предложен метод мультиагентного контроллера покрытия, объединяющий ряд взаимодополняющих механизмов, направленных на повышение стабильности обучения и эффективность координации агентов в задаче планирования пути покрытия. Эксперименты показали, что каждый компонент метода вносит существенный вклад в улучшение динамики покрытия, снижение дисперсии градиента и повышение устойчивости стратегий, а их совместное применение обеспечивает максимальную полноту и скорость покрытия области. Проведенный абляционный анализ подтверждает, что сочетание адаптивной ширины интервала обрезки, шлюзования модулированных преимуществ, контрфактического базиса и механизма многоголового самовнимания формирует эффективную архитектурную конфигурацию для решаемой задачи. Полученные результаты демонстрируют перспективность MACC для применения в реальных мультиагентных системах, требующих надежной координации и минимизации времени выполнения задач покрытия.

Исследование выполнено в рамках реализации гранта ФГБОУ ВО «Псковский государственный университет» по мероприятию «Выполнение прикладных научных и научно-методических исследований (проектов) молодыми учеными по приоритетным направлениям Программы развития университета на 2025-2036 годы» (приказ № 0905-001 от 05.09.2025).

Литература

1. Гао Т. Роботизированное обучение по демонстрациям с использованием диффузионной модели и алгоритмов обучения с

подкреплением // Инженерный вестник Дона, 2025, № 3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9945/.

2. Кузьмин Р. С., Алексеев И. В., Туманян М. М., Кемпи Е. А., Семенов Р. А., Лева Д. С., Рего Г. Э. Разработка системы для оптимизации планирования дислокации мест рубок на примере Республики Карелия // Инженерный вестник Дона, 2025, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10061/.

3. Абузяров А. А., Макаров А. А. Сравнение алгоритмов MCTS, MCDDQ, MCDDQ-SA, Greedy в рамках задачи параллельного планирования загрузки машин на производстве // Инженерный вестник Дона, 2025, № 6. URL: ivdon.ru/ru/magazine/archive/n6y2025/10139/.

4. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal Policy Optimization Algorithms // ArXiv. 2017. pp. 1–12.

5. Schulman J., Levine S., Abbeel P., Jordan M., Moritz P. Trust Region Policy Optimization // Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR W&CP, 2015. Vol. 37. pp. 1889–1897.

6. Foerster J. N., Farquhar G., Afouras T., Nardelli N., Whiteson S. Counterfactual Multi-Agent Policy Gradients // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018. Vol. 32, No. 1. pp. 2974–2982.

7. Vinyals O., Babuschkin I., Czarnecki W. M. [et al.] The StarCraft II Multi-Agent Challenge // ArXiv. 2017. pp. 1–20.

8. Das A., Gervet T., Romoff J., Batra D., Parikh D., Rabbat M., Pineau J. TarMAC: Targeted Multi-Agent Communication // Proceedings of the 36th International Conference on Machine Learning. Long Beach: Curran Associates, 2019. Vol. 97. pp. 1538–1546.

9. Engstrom L., Ilyas A., Santurkar S., Tsipras D., Janoos F., Rudolph L., Madry A. Implementation Matters in Deep RL: A Case Study on PPO and TRPO // ArXiv. 2020. pp. 1–14.

10. Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR W&CP, 2015. Vol. 37. pp. 448–456.

References

1. Gao T. Inzhenernyj vestnik Dona, 2025, No. 3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9945/.

2. Kuz'min R. S., Alekseev I. V., Tumanjan M. M., Kempi E. A., Semjonov R. A., Leva D. S., Rego G. Je. Inzhenernyj vestnik Dona, 2025, No. 5. URL: ivdon.ru/ru/magazine/archive/ n5y2025/10061/.

3. Abuzjarov A. A., Makarov A. A. Inzhenernyj vestnik Dona, 2025, No. 6. URL: ivdon.ru/ru/magazine/archive/n6y2025/10139/.

4. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. ArXiv, 2017, pp. 1–12.

5. Schulman J., Levine S., Abbeel P., Jordan M., Moritz P. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR W&CP, 2015, Vol. 37, pp. 1889–1897.

6. Foerster J. N., Farquhar G., Afouras T., Nardelli N., Whiteson S. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018, Vol. 32, No. 1, pp. 2974–2982.

7. Vinyals O., Babuschkin I., Czarnecki W. M. [et al.] ArXiv, 2017, pp. 1–20.

8. Das A., Gervet T., Romoff J., Batra D., Parikh D., Rabbat M., Pineau J. Proceedings of the 36th International Conference on Machine Learning. Long Beach: Curran Associates, 2019, Vol. 97, pp. 1538–1546.

9. Engstrom L., Ilyas A., Santurkar S., Tsipras D., Janoos F., Rudolph L., Madry A. ArXiv, 2020, pp. 1–14.

10. Ioffe S., Szegedy C. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR W&CP, 2015, Vol. 37, pp. 448–456.

Автор согласен на обработку и хранение персональных данных.

Дата поступления: 9.12.2025

Дата публикации: 22.02.2026