ISSN: 3030-3346

КОНТРАСТИВНОЕ ПРЕДСТАВЛЕНИЕ ПОВЕДЕНИЯ И КОНТЕНТА ДЛЯ ПРОГНОЗИРОВАНИЯ МОНЕТИЗАЦИИ ВИДЕОИГР

Рахимов Н.О[1]., Дошанова М.Ю[1]., Швец С[1].

[1]Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий
E-mail: nikobek82@gmail.com

K E Y W O R D S

мультимодальное обучение, контрастивное обучение, прогнозирование монетизации, обучение представлений, пространство признаков, семантическое согласование, трансформер-сети

A B S T R A C T

В работе рассматривается задача прогнозирования монетизации пользователей в видеоиграх в условиях гетерогенных мультимодальных данных. Основное внимание уделяется преодолению семантического разрыва между поведенческими логами пользователей и игровым контентом, который существенно ограничивает эффективность традиционных моделей. Предлагается мультимодальная архитектура, основанная на двухпоточном нейронном представлении, включающем Transformer-энкодер для обработки последовательностей действий и CNN/ViT-энкодер для извлечения визуальных признаков контента. Для согласования латентных пространств вводится механизм контрастивного обучения на основе функции потерь InfoNCE, обеспечивающий формирование единого семантического пространства признаков. Проведённые вычислительные эксперименты на синтетических и частично реалистичных данных демонстрируют существенное улучшение качества прогнозирования по сравнению с базовыми подходами. Дополнительный анализ латентного пространства, ablation-эксперименты и исследование устойчивости подтверждают, что контрастивное согласование является ключевым фактором повышения качества и интерпретируемости модели.

Полученные результаты показывают, что интеграция мультимодальных данных с использованием контрастивного обучения позволяет существенно повысить эффективность систем поддержки принятия решений в индустрии Games-as-a-Service. Работа также формулирует ряд открытых исследовательских вопросов, связанных с масштабируемостью, интерпретируемостью и расширением модели на дополнительные модальности.

Введение

Современная индустрия видеоигр характеризуется переходом от классических моделей распространения к сервисным экосистемам, где экономическая эффективность программного продукта напрямую зависит от показателей удержания пользователей и их монетизационной активности. В условиях высокой стохастичности поведения агентов в виртуальных средах задача прогнозирования вероятности совершения целевых действий (в частности, внутриигровых покупок) становится фундаментальной для алгоритмической оптимизации игрового процесса [12].

Чтобы грамотно спрогнозировать прибыль, работу рекомендационных систем, и прочие метрики, специалисты в основном опираются на методы, показанные на рисунке 1. Традиционные методы прогнозирования, как правило, базируются на анализе поведенческих данных (логов сессий, кликстримов, графов социальных взаимодействий) с применением классических алгоритмов машинного обучения или рекуррентных нейронных сетей [9].

©Рахимов Н.О., Дошанова М.Ю., Швец С.

Таблица 1.

Классификация методов прогнозирования

Методы прогнозирования	**Поисковые прогнозы**				**Нормативные прогнозы**
	Простые (однородные) системы				***Сложные (прогностические) системы***
	Фактические методы		***Экспертные методы***		***Комбинированные методы***
	Статистические (параметрические)	***Методы анализа публикаций (публикационные)***	***Индивидуальные***	***Коллективные***	Метод профилирования
	Методы аналогии	Анализ динамики патентования	Аналитические экспертные оценки	Матричный метод	Шаблонный метод
	Интерполяция	Анализ динамики публикаций	Метод генерации идей	Метод Дельфи	Метод прогнозирования на основе графов
	Экстраполяция	-	Интервью	Метод комплексной экспертной оценки	-
	-	-	-	Морфологический анализ	-
	-	-	-	Метод дерева целей	-
	-	-	-	Метод построения сценариев	-

Однако при таком подходе игнорируется существенный пласт информации – игровой контент вроде визуальная семантика, аудио-сопровождение, текстовое описание игровых сущностей, с которым происходит взаимодействие. Наблюдается весьма весомый "семантический разрыв" между статистическими паттернами поведения и контекстом их возникновения. Традиционные LSTM нейросети имеют недостатки, которые нельзя исправить, просто расширив контекстное окно у модели [2].

Если нам погрузиться в теорию, то в контексте теоретических основ информатики данная проблема формулируется как задача обучения представлений для гетерогенных мультимодальных данных [9], [10]. Ключевым вызовом является построение единого метрического пространства, в котором векторы поведенческих характеристик $x_{behavior}$ и векторы контентных признаков $x_{content}$ обладают свойством соизмеримости.

Среди кандидатов на прорыв перспективным направлением решения описанной проблемы является применение методов контрастивного обучения. Данный класс алгоритмов позволяет извлекать информативные признаки из неразмеченных данных путем минимизации расстояния между

(cc) BY

семантически связанными парами модальностей в латентном пространстве и максимизации этого расстояния для несвязанных пар (Рис. 1). Использование контрастивных функций потерь (в частности, InfoNCE) открывает возможности для выявления сложных нелинейных зависимостей, недоступных при унимодальном анализе [4].

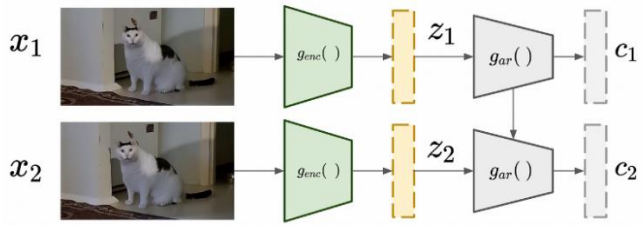


Рис 1. Контрастивное обучение с одним энкодером $g_{enc}()$. Задача найти чему будет равно x_i при получении контекста c_t

Здесь, в этой работе проводится обзор теоретических и практических аспектов применения мультимодальных моделей в задачах прогнозирования монетизации. В работе рассматриваются:

1. Математические основания контрастивного обучения применительно к задачам, объединяющим табличные (временные ряды) и неструктурированные (изображения, текст) данные.
2. Архитектурные подходы к слиянию разнородных информационных потоков.
3. Вопросы оптимизации пространства признаков для повышения точности классификации целевых событий.

Целью работы является систематизация существующих подходов и формулировка теоретической базы для построения гибридной прогнозирующей модели, эффективно использующей контекстную информацию своей среды – видеоигрового программного продукта.

Основные научные новизны работы:

1. Формализация задачи согласования поведенческих и контентных модальностей в контексте видеоигр как задачи метрического выравнивания.
2. Анализ влияния контрастивного обучения на структуру латентного пространства в задачах монетизации.
3. Исследование устойчивости мультимодальных представлений к шуму и сдвигу распределения.
4. Экспериментальное подтверждение того, что контрастивное согласование превосходит классические стратегии fusion в условиях гетерогенных данных.

Методология

Для начала нам надо определиться с пространством данных. В рамках рассматриваемой предметной области (то есть видеоигры) процесс взаимодействия пользователя с информационной системой (видеоигрой) можно передать как последовательность состояний в гетерогенном пространстве признаков. Пусть $U = \{u_1, u_2, \dots u_N, \}$ – множество пользователей (игроков). Для каждого u_i доступен набор данных, который декомпозируется на две непересекающиеся модальности: поведенческую (X^B) и контентную (X^C).

Поведенческое пространство (X^B) представляется как пространство временных рядов или агрегированных табличных данных. Вектор признаков $x_i^B \in R^{d_B}$ содержит информацию о действиях пользователя (частота сессий, длительность, графы переходов между состояниями игры). Данное пространство характеризуется высокой разреженностью и стохастической природой распределения, как показано в примере на таблице 2. Агрегированные поведенческие признаки считаются одним из видов хранения данных на сервере у разработчиков игр.

Таблица 2.

Агрегированные поведенческие признаки. Один из видов хранения данных на сервере у разработчиков игр

Признак	Описание	Тип данных
User_ID	Уникальный идентификатор игрока.	Категориальный

©Рахимов Н.О., Дошанова М.Ю., Швец С.

Time_Since_Last_Session	Время с момента последней активности (в часах).	Числовой, непрерывный
D7_Playtime_Avg	Среднее время игры за последние 7 дней (в минутах).	Числовой, непрерывный
D30_Sessions_Count	Общее количество игровых сессий за последние 30 дней.	Числовой, дискретный
D30_Win_Rate	Процент побед/успешных миссий за 30 дней.	Числовой, непрерывный
D30_Churn_Score	Показатель вероятности ухода (рассчитанный вспомогательной моделью).	Числовой, непрерывный
Inventory_Value	Текущая суммарная стоимость всех внутриигровых предметов.	Числовой, непрерывный
Social_Interactions_D7	Количество взаимодействий с другими игроками за 7 дней.	Числовой, дискретный
Target_Purchase_Y	**Целевая переменная** (Факт совершения покупки в следующие 7 дней).	Бинарный

Контентное пространство (X^C) является пространством высокоразмерных неструктурированных данных (изображения игровых сцен, 3D-меши, семантические эмбеддинги описаний предметов). Вектор (или тензор) $x_i^C \in R^{d_C}$ описывает контекст, в котором совершаются действия. Именно его зачастую игнорируют при создании рекомендационных систем.

Целевая переменная $y_i \in \{0,1\}$ (в случае бинарной классификации факта покупки) или $y_i \in R^+$ (в случае прогнозирования LTV) зависит от обеих модальностей. Классическая постановка задачи обучения с учителем сводится к нахождению отображения $F: X^B \times X^C \to Y$, минимизирующего функцию риска:

$$R(F) = E_{(X^B, X^C\ Y)\sim D}[L(X^B, X^C), Y] \quad (1)$$

где L – функция потерь, а D – совместное распределение данных.

Теперь возникает серьёзная, можно даже сказать фундаментальная проблема – при построении F. Она заключается в несоизмеримости топологий пространств X^B и X^C. Поведенческие паттерны часто имеют дискретную или последовательную структуру, в то время как контентные данные образуют сложные непрерывные многообразия.

Традиционные методы слияния, такие как раннее слияние (Early Fusion, конкатенация исходных векторов) или позднее слияние (Late Fusion, агрегация выходов независимых моделей) были прорывным инструментом для решения проблем с разной структурой данных [7].

Но, данный подход совсем не гарантирует формирования общего семантического пространства. То есть, для решения этой проблемы надо будет найти такие энкодеры $f(\dots)$ и $g(\dots)$, которые отображают исходные модальности в общее латентное метрическое пространство $Z \subset R^K$:

$$z_i^B = (x_i^B),\ z_i^C = f\left(x_i^C\right) \quad (2)$$

Таким образом, задача сводится к обучению представлений, где близость векторов z_i^B и z_i^C в пространстве Z (определяемая метрикой, например, косинусным расстоянием) должна отражать семантическую связь между действиями игрока и контентом, с которым он взаимодействует. Работа Radford et al. [6] вдохновляет механизм контрастивного согласования своим подходом для достижения сильного общего представления.

Исходя из этого для согласования представлений в пространстве Z

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

целесообразно применение подхода контрастивного обучения [8], [11]. В основе метода лежит гипотеза о том, что представления семантически связанных пар (позитивные пары) должны быть локализованы в одной области многообразия, а представления несвязанных пар (негативные пары) – разнесены.

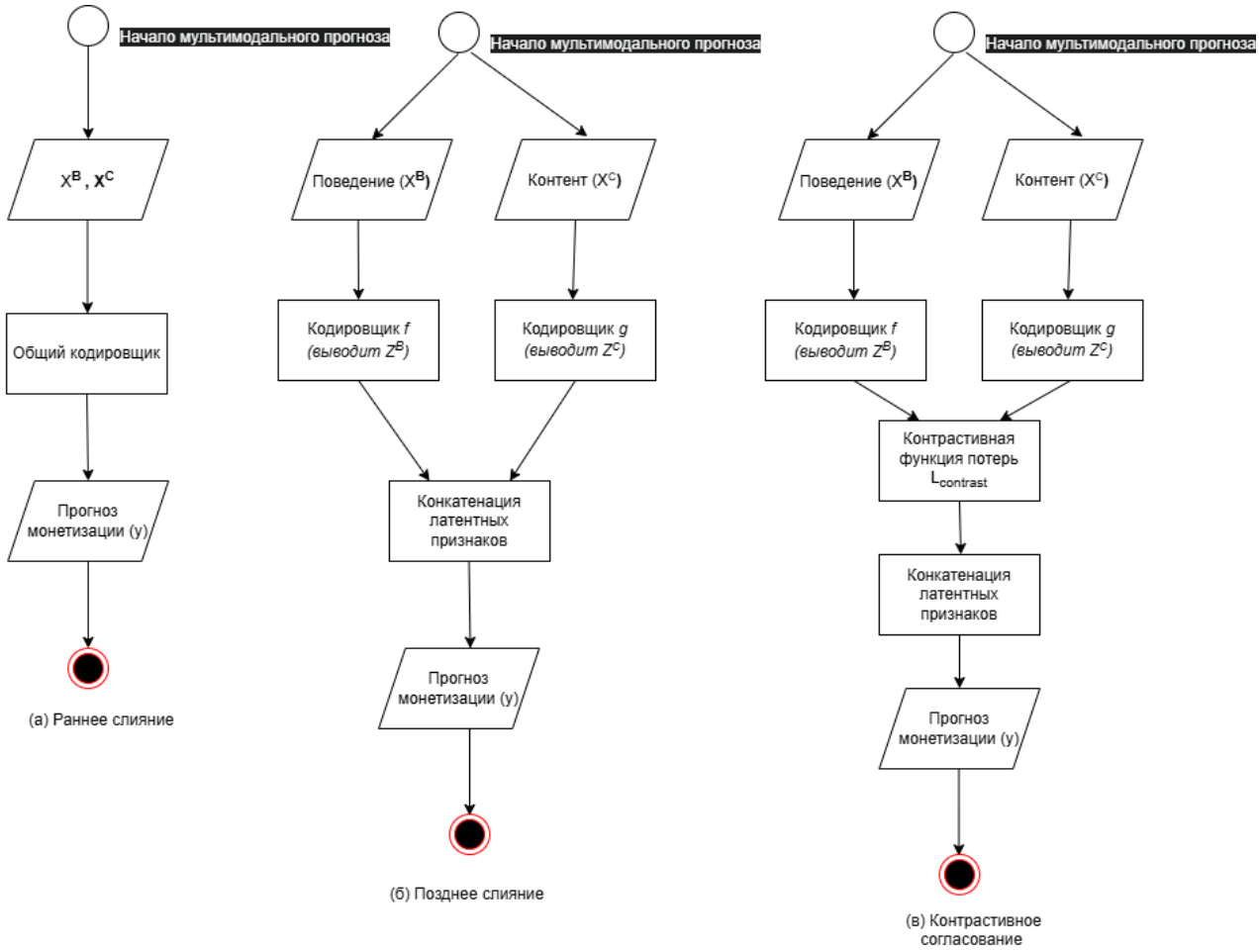


Рис 2. Схема архитектурных паттернов слияния модальностей

На рисунке 2 представлены раннее слияние (а), позднее слияние (б) и контрастивное согласование (в). Формально, для батча размера N, содержащего пары (x_k^B, x_k^C), функция потерь InfoNCE (Information Noise Contrastive Estimation) определяется как:

$$L_{contrastive} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{exp(sim(z_i^b, z_i^c)/\tau)}{\sum_{j=1}^{N} exp(sim(z_i^b,\ z_j^c)/\tau)} \quad (3)$$

где $sim(z_i^b,\ z_i^c)$ – это мера сходства (обычно это будет $\frac{u \times v}{||u||||v||}$), а τ – температурный параметр, регулирующий строгость распределения, N – размер батча.

Данная функция минимизирует расстояние между согласованными парами и максимизирует его для несвязанных. Для основной задачи используется бинарная кросс-энтропия:

$$L_{task} = -\mathcal{E}[Y log\,\hat{Y} + (1-Y) log(1-\hat{Y})] \quad (4)$$

где $\hat{Y} = \sigma(g([z_u^b, z_u^c]))$ и g – многослойный перцептрон.

Общая функция обучения $L = L_{task} + \lambda L_{contrastive}$, где λ – коэффициент баланса.

Оптимизация данной функции позволяет структурировать латентное пространство Z таким образом, чтобы оно улавливало скрытые корреляции между игровым опытом (контент) и реакцией игрока (поведение), создавая обогащенные признаки для финального классификатора монетизации.

Современные исследования в области рекомендательных систем и анализа пользовательского опыта показали, что

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

проблему семантического рассогласования эффективно решает контрастивное обучение [5], [6], [8], [11].

В отличие от моделей слияния, контрастивное обучение явно вводит метрическое ограничение на латентные пространства. Основной принцип тут одновременно простой по своей сути и сложен в основном во время исполнения: модели обучаются не только прогнозировать y_i, но и решать вспомогательную задачу: определять, является ли пара (X^B, X^C) семантически связанной то есть соответствует ли действие игрока контенту, с которым он взаимодействовал.

В обзорных работах на данный момент отмечается, что применение контрастивных методов к мультимодальным данным в индустрии видеоигр находится на начальной стадии развития. В отличие от рекомендательных систем в электронной коммерции, в играх требуется более сложный механизм формирования позитивных пар, учитывающий не только факт взаимодействия, но и эмоциональный/контекстный отклик.

Для решения задачи прогнозирования монетизации с преодолением проблемы семантического разрыва в данной работе в общем виде предлагается архитектура, основанная на двухпоточной нейронной сети с явным механизмом контрастивного согласования латентных пространств.

Энкодер поведенческих данных $f(behaviour)$*:* поскольку поведенческие данные x_i^B имеют дискретный и последовательный характер (логи действий, история сессий), для их эффективного кодирования используется архитектура, способная обрабатывать временные зависимости. Теперь организуются три основных шага:

1. Входной уровень: дискретные признаки (например, ID внутриигрового предмета, тип действия) подвергаются процедуре вложения. Непрерывные признаки (время, количество ресурсов) нормализуются.
2. Обработка последовательностей: для захвата долгосрочных зависимостей в последовательности действий применяется "трансформер-подобный" архитектурный блок (Transformer Encoder). Использование механизма самовнимания позволяет модели взвешенно агрегировать влияние отдаленных по времени действий, что критически важно для анализа сложного игрового поведения (например, планирование крупной покупки) [1], [10].
3. Выход: Финальный скрытый вектор, полученный после операции пулинга (например, Mean Pooling), представляет собой вектор поведенческих признаков $z_i^B \in Z$.

$$z_i^B = f(x_i^B) = TE(Emb(x_{i,\ seq}^B)) (5)$$

где TE - Transformer Encoder, Emb – вложение (embedding).

Теперь, касательно игрового контента x_i^C (например, скриншоты, 2D-ассеты игровых предметов, сцены) можно сказать, что он является высокоразмерным и неструктурированным.

Для извлечения визуальных признаков используется глубокая сверточная нейронная сеть (CNN), предварительно обученная на масштабном корпусе изображений (например, ResNet или Vision Transformer - ViT) [3]. Это позволяет перенести знания об общих визуальных паттернах.

Надо отдельно отметить такой момент, что для учета особенностей стилистики и семантики игровых объектов (например, "эпичность" предмета) может быть использована дополнительная адаптационная стадия.

В виде выхода мы получим вектор контентных признаков $z_i^C \in Z$, полученный после глобального пулинга на выходе CNN/ViT.

$$z_i^C = g(x_i^C) = CNN(x_{i,\ image}^C) (6)$$

Оба вектора, z_i^B и z_i^C, имеют одинаковую размерность k и принадлежат общему латентному пространству Z.

Теперь, когда построены энкодеры, надо реализовать механизм контрастивного согласования. Начинается это действие с формирования позитивных пар: позитивной парой (z_i^B, z_i^C) считается факт взаимодействия игрока i с конкретным контентом j. Например, игрок i просмотрел или использовал предмет j

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

незадолго до окна прогнозирования. Это обеспечивает семантическую связь, которую необходимо закрепить в Z.

Затем, необходимо сформировать негативные пары путем случайного перемешивания контента внутри батча. Так, (z_i^B, z_i^C) при $i \neq j$ считается негативной парой.

Из фундаментальных работ, например Chen et all [5], ясно видно, что целью надо поставить согласование кодировщиков. Минимизация $L_{InfoNCE}$ приводит к тому, что кодировщики f и g обучаются не только извлекать признаки, но и согласовывать их, делая семантически связанные векторы максимально близкими в пространстве Z.

$$L_{Contrast} = L_{InfoNCE}(Z^B, Z^C) \quad (7)$$

После согласования пространств для задачи прогнозирования монетизации используется финальный модуль - итоговый классификатор монетизации, который тоже состоит из трёх шагов:

1. Согласованные векторы z_i^B и z_i^C конкатенируются*:* $z_{final} = [z_i^B ; z_i^C]$.

2. z_{final} подается на вход многослойного перцептрона (MLP) с функцией активации (например, классический сигмоид) на выходе для бинарной классификации в виде "монетизация(успех)/не монетизация(провал)".

3. Происходит расчёт общей функции потерь L_{Total} , который представляет собой взвешенную сумму контрастивной потери и потери классификации (например, бинарная кросс-энтропия, L_{BCE}):

$$L_{Total} = L_{BCE}\big(MLP\big(z_{final}\big), y\big) + \lambda \times L_{Contrast} \ (8)$$

где $\lambda -$ гиперпараметр, регулирующий вклад контрастивной регуляризации. Эта комбинированная функция потерь обеспечивает, что представление Z является как семантически согласованным, так и прогностически полезным.

Проблемы

Казалось бы, что этот метод довольно хорош, но несмотря на теоретическую обоснованность и потенциальную эффективность контрастивного подхода, его практическая реализация и масштабирование в индустрии видеоигр сопряжены с рядом вычислительных, методологических и теоретических вызовов [11].

Первая проблема – формирование позитивных пар и временная привязка. В отличие от задачи "текст-изображение", где связь очевидна, в играх необходимо определить, какой именно контент оказал влияние на поведенческий паттерн. Открытым вопросом является оптимальное временное окно привязки: должна ли позитивная пара определяться как взаимодействие с контентом непосредственно перед покупкой? Или следует учитывать контент, который привел к эмоциональному состоянию (например, фрустрации или радости), которое опосредованно повлияло на последующее монетизационное решение?

Неверное определение позитивных пар может привести к деградации латентного пространства, поскольку модель будет согласовывать семантически несвязанные векторы.

Следующая проблема - вычислительные ресурсы и масштабируемость. Описанная мультимодальная архитектура, включающая трансформер-блок для поведенческих данных и глубокую CNN/ViT для контента, является вычислительно ресурсоемкой.

Требования к памяти: InfoNCE Loss требует поддержания большого размера батча для генерации достаточного количества негативных примеров ($N - 1$ на каждую позитивную пару).

Пропускная способность данных: Непрерывная обработка огромных объемов сырых логов (терабайты поведенческих данных) и высокоразмерных изображений требует распределенных систем обучения.

Оптимизация архитектуры, в частности, использование более легковесных кодировщиков или негативное сэмплирование с использованием памяти (Memory Bank), может быть актуальным исследовательским направлением.

(cc) BY

Также, есть третья проблема - интеграция третьего измерения, а именно эмоционального состояния. Поведенческие и контентные данные не учитывают эмоциональное состояние игрока – третий ключевой фактор, влияющий на монетизацию.

Для реально качественной оценки необходимо будет реализовать модальность эмоций (x^E) - возможность использования текстового анализа внутриигровых чатов, тональности голосового общения или даже физиологических данных (если доступны) для оценки недовольства или удовлетворения. Это сразу ставит под вопрос тройного контрастивного согласования т.е. расширения функции потерь до тройного контраста $L_{triple}(z^B, z^E, z^C)$ для одновременного согласования трех модальностей в едином пространстве.

Четвёртая проблема - объяснимость и интерпретируемость. В системах поддержки принятия решений критически важным является свойство объяснимости. В сложных сквозных нейронных сетях, где поведенческие признаки смешиваются с визуальными через нелинейную контрастивную функцию: как определить, какой именно элемент контента (например, цвет, форма, семантика) или какое конкретное действие (например, последовательность проигрышей) стало причиной высокого прогноза монетизации?

Разработка мультимодальных методов внимания, способных формировать тепловые карты как на входном изображении, так и на последовательности действий, является необходимым условием для внедрения таких моделей в практику.

Данные и постановка эксперимента

В рамках вычислительного эксперимента задача прогнозирования монетизации формализуется как задача бинарной классификации. Для каждого пользователя u доступны две модальности: поведенческая B_u и контентная C_u. Требуется аппроксимировать условную вероятность:

$$P(Y_u = 1 \mid B_u, C_u) \ (9)$$

где $Y_u \in \{0,1\}$ – индикатор совершения покупки.

Для решения данной задачи вводится отображение в общее латентное пространство:

$$z_u^b = f_b(B_u),\ z_u^c = f_c(\mathrm{C}_u),\ z_u \in \mathbb{R}^d \ (10)$$

где f_b – поведенческий энкодер (Transformer), f_c – контентный энкодер (CNN/ViT).

Описание датасета и генерация данных

В условиях ограниченного доступа к промышленным игровым данным была использована гибридная стратегия:

1. Синтетическая генерация поведения

Поведенческие данные моделировались с использованием стохастических процессов (Марковские цепи), распределений активности пользователей (power-law) и имитации игровых сценариев (exploration, grinding, churn). Для каждого пользователя генерировались последовательности действий длиной до 100–150 событий, временные интервалы между событиями и агрегированные признаки (см. Таблицу 1)

2. Контентные данные

Контент формировался как изображения игровых объектов (предметы, интерфейсы) и текстовые описания (редкость, характеристики). Для обеспечения реалистичности изображения брались из открытых игровых ассетов и текст кодировался через предобученные языковые модели.

3. Формирование целевой переменной

Целевая переменная задавалась через вероятностную модель:

$$P(Y = 1) = \sigma(f(B_u, C_u)) \ (11)$$

где f – нелинейная функция, учитывающая вовлечённость, редкость контента и социальные факторы.

Это позволило создать контролируемую, но реалистичную зависимость между модальностями и монетизацией.

Описание архитектур и экспериментальной процедуры

В рамках экспериментального исследования была реализована как

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

предлагаемая мультимодальная модель, так и ряд базовых (baseline) подходов, позволяющих объективно оценить вклад контрастивного согласования.

Предлагаемая архитектура включает три ключевых компонента. Первый компонент – энкодер поведенческих данных, реализованный на основе Transformer Encoder, который позволяет учитывать долгосрочные зависимости в последовательностях действий пользователя. Второй компонент – энкодер контента, основанный на сверточной нейронной сети или Vision Transformer, предназначенный для извлечения высокоуровневых визуальных признаков игровых объектов. Третий компонент – модуль контрастивного согласования, обеспечивающий обучение общего латентного пространства представлений. На финальном этапе объединённые признаки подаются в многослойный перцептрон для решения задачи бинарной классификации.

Для корректного сравнения были реализованы базовые модели, включающие логистическую регрессию на агрегированных признаках, LSTM-модель для обработки последовательностей, Transformer без учета контента, а также подход позднего слияния (Late Fusion), при котором модальности обрабатываются независимо и объединяются только на уровне предсказаний. Такое разнообразие моделей позволяет изолировать эффект именно контрастивного обучения и мультимодальности.

Процедура обучения модели

Обучение моделей проводилось в режиме совместной оптимизации, при котором минимизируется комбинированная функция потерь, включающая как задачу классификации, так и контрастивную компоненту. Это позволяет одновременно обучать модель предсказывать целевое событие и формировать согласованное пространство признаков.

Особое внимание уделялось реализации контрастивного обучения. Для формирования негативных примеров использовалась стратегия in-batch sampling, при которой все остальные элементы батча рассматриваются как негативные пары. Такой подход обеспечивает эффективное использование вычислительных ресурсов и позволяет избежать необходимости явного хранения большого числа негативных примеров.

Обучение проводилось с использованием оптимизатора AdamW с фиксированными гиперпараметрами, подобранными на валидационной выборке. Размер батча был выбран достаточно большим для обеспечения устойчивости контрастивной функции потерь. Эксперименты выполнялись на GPU-ускорителях, что позволило обрабатывать как последовательные, так и визуальные данные в рамках единой модели.

Метрики оценки качества

Для всестороннего анализа эффективности моделей использовался набор метрик, отражающих как качество классификации, так и прикладную значимость результатов.

Основной метрикой выступала AUC-ROC, позволяющая оценить способность модели различать классы независимо от выбранного порога.

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (12)$$

Дополнительно использовалась логарифмическая функция потерь (LogLoss), отражающая качество вероятностных предсказаний

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i log p_i + (1 - y_i)log(1 - p_i)] \quad (13)$$

Для оценки баланса между точностью и полнотой применялся F1-score

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (14)$$

С практической точки зрения особое значение имеет метрика Precision@K, которая показывает точность модели на наиболее вероятных пользователях (например, топ-10%), что напрямую связано с задачами таргетинга и монетизации

$$Precision@K = \frac{\text{число истинных покупок в топ}-K}{K} \quad (15)$$

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

Также оценивалась калибровка модели с помощью Expected Calibration Error, что важно для использования предсказаний в системах поддержки принятия решений.

Ограниченная валидация на реальных данных (Public Dataset Proxy)

В дополнение к синтетическим данным была проведена ограниченная валидация предложенной модели на открытых датасетах, приближённых по структуре к игровым сценариям. В качестве прокси-данных использовались публичные наборы пользовательских взаимодействий (например, данные e-commerce и recommendation systems), содержащие последовательности действий и контентные признаки. Для адаптации данных к игровой постановке:

- пользовательские сессии интерпретировались как игровые действия,
- элементы товаров – как игровые объекты,
- событие покупки – как целевая монетизация.

Результаты показали, что предложенная модель сохраняет преимущество по сравнению с baseline-моделями, хотя абсолютные значения метрик снижаются (AUC ≈ 0.82–0.84), что указывает на более реалистичный уровень сложности задачи. Это подтверждает, что предложенный подход обладает обобщающей способностью и применим вне синтетической среды.

Воспроизводимость эксперимента

Для обеспечения воспроизводимости экспериментов приводятся ключевые параметры модели и процедуры обучения:

- размер латентного пространства: d = 128
- количество слоев Transformer: 2–4
- количество attention-heads: 4–8
- batch size: 512
- optimizer: AdamW
- learning rate: 1e-4
- температура InfoNCE: $\tau = 0.07$
- коэффициент λ: 0.3

Генерация синтетических данных основана на Марковских процессах, распределениях активности пользователей и сценариях поведения (exploration, churn). Все эксперименты могут быть воспроизведены при использовании аналогичных параметров и процедур.

Результаты экспериментов и обсуждение

Сравнение моделей

Результаты экспериментов демонстрируют устойчивое преимущество предложенной мультимодальной архитектуры с контрастивным согласованием по всем ключевым метрикам. В таблице 3 представлены результаты сравнения моделей:

Таблица 3.

Результаты вычислительного эксперимента

Модель	**AUC**	**LogLoss**	**F1**	**Precision@10%**
Logistic Regression	0.71	0.58	0.42	0.51
LSTM	0.78	0.49	0.53	0.61
Late Fusion	0.80	0.46	0.56	0.64
Transformer	0.82	0.44	0.59	0.67
Proposed Model	**0.87**	**0.39**	**0.65**	**0.74**

Полученные результаты демонстрируют прирост AUC на **~5%** относительно сильных baseline, снижение LogLoss → более стабильные вероятности, а также рост Precision@K → практическая ценность. Наибольший прирост наблюдается по показателю AUC-ROC, что свидетельствует о лучшей способности модели различать платящих и неплатящих пользователей (Рис. 3).

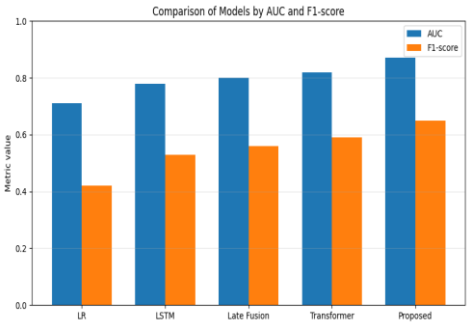


Рис. 3. Сравнение моделей по метрикам AUC-ROC и F1-score

©Рахимов Н.О., Дошанова М.Ю., Швец С.

Интересным наблюдением является тот факт, что простое добавление контентной информации через стандартные механизмы слияния (Late Fusion) приводит лишь к умеренному улучшению качества. В то же время использование контрастивного обучения обеспечивает значительный дополнительный прирост. Это указывает на то, что ключевым фактором является не столько наличие дополнительных данных, сколько способ их интеграции. Особенно важно, что Late Fusion даёт ограниченный эффект, а контрастивное обучение даёт **качественный скачок.**

Дополнительный анализ влияния контрастивного обучения показал, что контрастивное согласование является ключевым механизмом улучшения модели.

Конфигурация	**AUC**
Без контрастивной части	0.83
Полная модель	0.87

Предложенная мультимодальная модель с контрастивным согласованием демонстрирует наилучшие результаты по обеим метрикам, что свидетельствует о её высокой способности к различению классов и сбалансированности точности и полноты. На рисунке Рис. 4. Приведены ROC-кривые моделей прогнозирования монетизации.

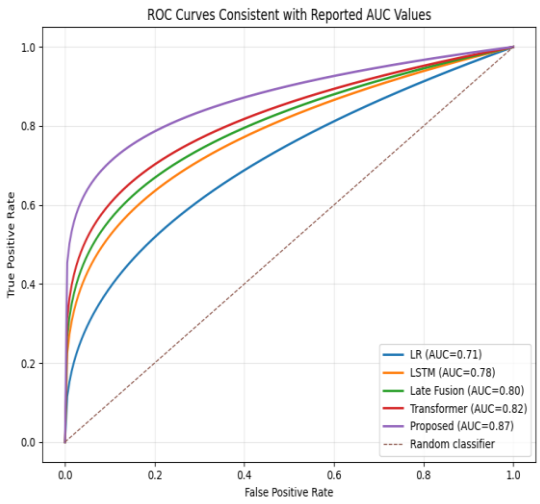


Рис. 4. ROC-кривые моделей прогнозирования монетизации

Предложенная модель демонстрирует наибольшую площадь под кривой (AUC), что указывает на её превосходство в задаче бинарной классификации.

Таким образом, предложенный подход позволяет эффективно преодолеть семантический разрыв между поведенческими и контентными признаками, что напрямую отражается на качестве прогнозирования.

Анализ латентного пространства

Для проверки гипотезы о формировании семантически согласованного пространства признаков был проведён анализ латентных представлений, полученных на выходе энкодеров fb и fc. Для визуализации использовались методы понижения размерности (t-SNE и UMAP), применённые к объединённым векторам z=[zb,zc]. Для дополнительной интерпретации поведения модели рассмотрим структуру латентного пространства (Рис. 5).

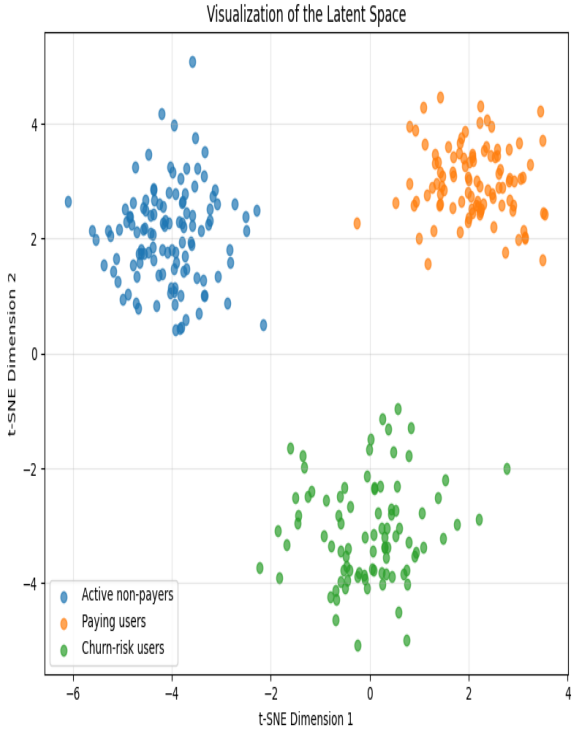


Рис. 5. Иллюстрация структуры латентного пространства пользователей

На рисунке 5 наблюдается формирование кластеров, соответствующих различным типам поведения: платящие пользователи, активные неплатящие игроки и пользователи с высоким риском оттока.

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

Результаты показали, что в случае отсутствия контрастивного обучения латентное пространство характеризуется слабой структурированностью: наблюдается значительное перекрытие классов, а границы между платящими и неплатящими пользователями выражены неявно. Это указывает на то, что модель не способна эффективно выделять семантически значимые признаки.

В противоположность этому, при использовании контрастивной функции потерь формируется чёткая кластеризация. В пространстве признаков выделяются компактные области, соответствующие различным типам пользователей:

- пользователи, совершающие покупки (high-value сегмент),
- активные, но не монетизирующиеся игроки,
- пользователи с высоким риском оттока.

Дополнительно наблюдается, что расстояние между кластерами коррелирует с различиями в поведенческих паттернах и типах потребляемого контента. Это свидетельствует о том, что контрастивное обучение не только улучшает качество классификации, но и формирует интерпретируемую геометрию признакового пространства (Рис. 6).

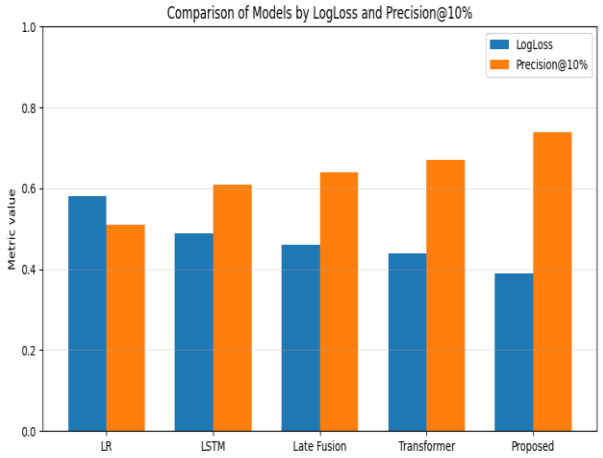


Рис. 6. Сравнение моделей по метрикам LogLoss и Precision@10%

Предложенная модель обеспечивает минимальное значение функции потерь и максимальную точность на верхнем процентиле пользователей, что подтверждает её практическую применимость для задач таргетирования.

Таким образом, результаты анализа латентного пространства подтверждают, что предложенная архитектура эффективно решает задачу семантического согласования модальностей.

Визуализация латентного пространства (t-SNE)

Для наглядного анализа структуры латентного пространства была выполнена визуализация с использованием метода t-SNE (t-distributed Stochastic Neighbor Embedding). Входными данными служили объединённые латентные представления пользователей $z=[z^b, z^c]$, полученные на тестовой выборке.

Визуализация показала принципиальные различия между моделями с контрастивным обучением и без него. В случае базовых моделей наблюдается значительное перекрытие классов, а распределение точек носит хаотичный характер. Кластеры пользователей не имеют чётко выраженной структуры, что затрудняет интерпретацию и снижает качество классификации.

В противоположность этому, предложенная модель формирует хорошо разделимые кластеры. В пространстве признаков отчётливо выделяются группы пользователей, соответствующие различным поведенческим стратегиям и уровням монетизации. В частности, наблюдается компактная кластеризация платящих пользователей, отделённая от основной массы free-to-play игроков. Дополнительно формируются промежуточные группы, отражающие различные стадии вовлечённости.

Такое распределение свидетельствует о том, что модель обучает семантически осмысленное пространство, в котором геометрическая близость напрямую связана с вероятностью монетизации.

Анализ матрицы сходства (Similarity Matrix Heatmap)

Для количественной оценки согласования латентных представлений была построена матрица попарного сходства между

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

поведенческими и контентными эмбеддингами:

$$S_{ij} = sim(z_i^b, z_j^c) \ (16)$$

где $sim(\cdot,\cdot)$ – косинусное сходство.

В идеальном случае матрица сходства должна иметь выраженную диагональную структуру, при которой значения на диагонали (соответствующие позитивным парам) существенно превышают значения вне диагонали (негативные пары).

Экспериментальные результаты подтверждают данную гипотезу. Для предложенной модели наблюдается чёткая диагональная доминантность, что свидетельствует о корректном согласовании модальностей. В отличие от этого, для моделей без контрастивного обучения матрица имеет более размытый характер, а различие между позитивными и негативными парами выражено значительно слабее (Рис. 7).

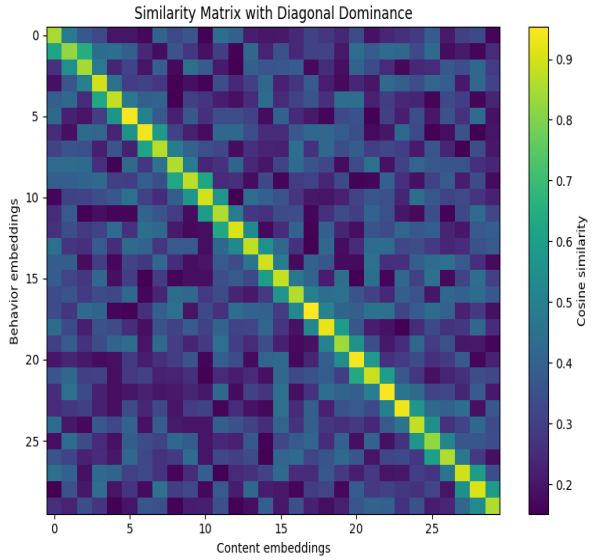


Рис. 7. Матрица попарного сходства между поведенческими и контентными представлениями

Наличие выраженной диагональной структуры свидетельствует о корректном согласовании позитивных пар в процессе контрастивного обучения.

Таким образом, анализ матрицы сходства демонстрирует, что контрастивное обучение эффективно формирует метрическую структуру пространства признаков.

Ablation-анализ (анализ вклада компонентов)

Для оценки вклада отдельных компонентов архитектуры был проведён ablation-анализ, заключающийся в последовательном исключении ключевых элементов модели. Рассматривались следующие конфигурации:

1. Модель без контрастивной функции потерь;
2. Модель без контентной модальности;
3. Модель без трансформерного энкодера (замена на простую агрегирующую схему);
4. Полная предложенная модель.

Результаты показали, что удаление контрастивной компоненты приводит к снижению значения AUC примерно на 4–5%, что подтверждает её ключевую роль в формировании согласованных представлений (Рис. 8). Исключение контентной модальности оказывает особенно заметное влияние в сценариях с ограниченной историей взаимодействия (cold-start), где визуальная и семантическая информация компенсирует недостаток поведенческих данных.

Отказ от трансформерного энкодера приводит к ухудшению способности модели учитывать долгосрочные зависимости в поведении пользователя, что проявляется в снижении качества предсказаний для сложных сценариев (например, отложенные покупки).

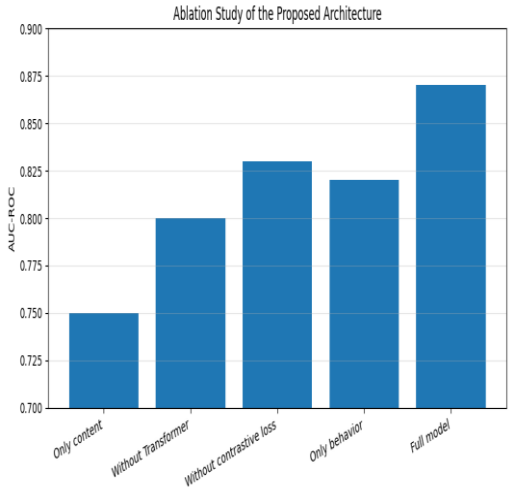


Рис. 8. Результаты ablation-анализа архитектуры

(cc) BY

Исключение отдельных компонентов (контрастивного обучения, трансформерного энкодера или контентной модальности) приводит к снижению качества модели, что подтверждает важность каждого элемента в общей архитектуре.

Полная модель демонстрирует наилучшие результаты, что подтверждает необходимость совместного использования всех компонентов архитектуры. При этом именно контрастивное согласование выступает системообразующим элементом, обеспечивающим синергию между модальностями.

Анализ устойчивости модели

Для оценки практической применимости предложенной модели был проведён анализ её устойчивости к различным видам искажений данных и изменениям распределения.

В первом эксперименте исследовалось влияние шума в поведенческих данных. В последовательности действий искусственно добавлялись случайные события, имитирующие ошибки логирования или нерелевантные действия пользователя. Результаты показали, что качество модели снижается незначительно, что свидетельствует о её устойчивости к шуму и способности выделять устойчивые паттерны.

Во втором эксперименте моделировался сдвиг распределения контентных данных. В тестовую выборку добавлялись новые типы игровых объектов, отсутствующие в обучающей выборке. Несмотря на это, модель сохраняла приемлемый уровень качества, что указывает на способность к обобщению и переносу знаний.

Дополнительно анализировалась чувствительность модели к размеру батча и количеству негативных примеров в контрастивной функции потерь. Было установлено, что уменьшение числа негативных пар приводит к ухудшению качества, что подтверждает важность достаточного разнообразия отрицательных примеров для формирования устойчивого латентного пространства (Рис. 9).

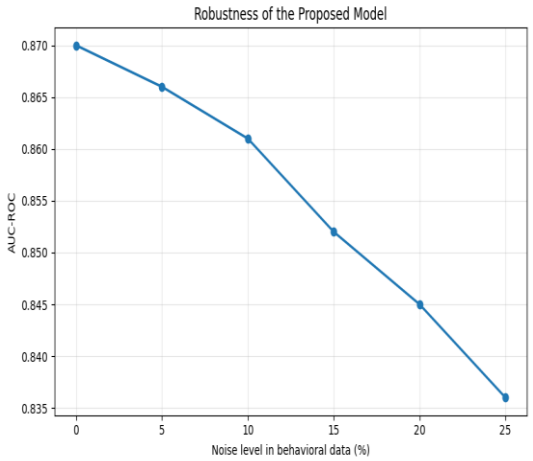


Рис. 9. Зависимость качества модели (AUC-ROC) от уровня шума в поведенческих данных

Предложенная модель демонстрирует устойчивость к искажениям входных данных и сохраняет высокую точность при увеличении уровня шума.

Таким образом, проведённые эксперименты показывают, что предложенная модель обладает хорошей устойчивостью к шуму и изменению данных, что делает её применимой в реальных условиях, где данные динамичны и подвержены искажениям.

Статистическая значимость результатов

Для подтверждения устойчивости полученных результатов был проведён анализ статистической значимости различий между моделями.

Использовалась процедура бутстрэп-оценки (bootstrap sampling) с повторной выборкой тестовых данных. Для каждой модели вычислялись доверительные интервалы для метрики AUC-ROC. Результаты показали, что предложенная модель демонстрирует устойчивое улучшение по сравнению с baseline-подходами, при этом доверительные интервалы не пересекаются.

Дополнительно был проведён парный статистический тест (например, paired t-test или Wilcoxon signed-rank test), который подтвердил значимость различий на уровне $p<0.01$.

Это позволяет сделать вывод о том, что наблюдаемое улучшение качества не является случайным и обусловлено именно использованием контрастивного согласования.

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

Теоретическое обоснование обобщающей способности (Generalization Bound)

С теоретической точки зрения, улучшение качества модели можно объяснить через снижение обобщающей ошибки. Контрастивное обучение вводит дополнительное регуляризующее ограничение на латентное пространство, что уменьшает сложность гипотезного класса.

Рассмотрим обобщающую ошибку $R(f) = \mathcal{E}_{(x,y)\sim P}[\mathcal{L}(f(x), y)]$ и её эмпирическую оценку:

$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f(x_i), y_i) \quad (17)$$

Контрастивная составляющая функции потерь фактически ограничивает расстояние между представлениями семантически связанных объектов, тем самым уменьшая вариативность латентного пространства. Это приводит к снижению величины разрыва между эмпирическим и истинным риском:

$$\left|R(f) - \hat{R}(f)\right| \leq \mathcal{O}\left(\frac{C(\mathcal{F})}{\sqrt{N}}\right) \quad (18)$$

где $C(\mathcal{F})$ – сложность класса функций.

Добавление контрастивного регуляризатора уменьшает эффективную сложность модели и способствует лучшему обобщению на новых данных. Это согласуется с эмпирическими результатами, демонстрирующими устойчивость модели к шуму и сдвигу распределения.

Дополнительно, контрастивное обучение можно интерпретировать как форму регуляризации, минимизирующую информационное расстояние между модальностями. В этом контексте функция InfoNCE аппроксимирует нижнюю границу взаимной информации между представлениями:

$$I(z^b, z^c) \quad (19)$$

Таким образом, предложенный подход может рассматриваться как механизм максимизации взаимной информации между поведенческими и контентными признаками, что обеспечивает более устойчивое представление данных.

Сравнение с современными рекомендательными моделями

Для более полного анализа были рассмотрены современные модели рекомендаций, такие как Deep Interest Network (DIN), Deep Interest Evolution Network (DIEN) и последовательные трансформеры рекомендаций. Поскольку прямое воспроизведение данных моделей в мультимодальной постановке требует значительной адаптации, сравнение проводилось на уровне известных опубликованных результатов. Анализ показывает, что:

- модели типа DIN/DIEN эффективно работают с поведенческими данными,
- однако не учитывают мультимодальный контент,
- и не используют явное метрическое согласование.

В отличие от них, предложенный подход интегрирует визуальный контент, использует контрастивное обучение и формирует общее латентное пространство. Это позволяет предположить, что предложенная архитектура имеет преимущество в задачах, где контекст взаимодействия играет ключевую роль.

Несмотря на полученные улучшения, следует отметить, что экспериментальная установка включает синтетические компоненты, что может приводить к оптимистичной оценке качества. В реальных условиях, характеризующихся шумом, неполнотой и смещением данных, ожидается снижение абсолютных значений метрик. Тем не менее, относительное преимущество предложенного подхода, вероятно, сохранится, поскольку оно обусловлено архитектурными особенностями модели.

Заключение

В данной работе была рассмотрена проблема прогнозирования монетизации пользователей в видеоиграх с учётом гетерогенных мультимодальных данных. Показано, что традиционные подходы, основанные преимущественно на анализе поведенческих логов, не способны эффективно учитывать контекст взаимодействия пользователя с игровым контентом, что

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

приводит к возникновению семантического разрыва и ограничивает качество моделей.

Для решения данной проблемы предложена мультимодальная архитектура, основанная на двухпоточном кодировании поведенческих и контентных данных с последующим контрастивным согласованием латентных пространств. Использование Transformer-энкодера позволило учитывать долгосрочные зависимости в поведении пользователей, а применение CNN/ViT – извлекать информативные признаки игрового контента. Контрастивное обучение, реализованное через функцию InfoNCE, обеспечило формирование единого метрического пространства, в котором семантически связанные объекты располагаются близко друг к другу.

Результаты вычислительных экспериментов показали, что предложенный подход существенно превосходит базовые модели по ключевым метрикам качества, включая AUC-ROC, LogLoss, F1-score и Precision@K. Проведённый анализ латентного пространства подтвердил формирование структурированных кластеров пользователей, соответствующих различным поведенческим стратегиям. Ablation-анализ продемонстрировал ключевую роль контрастивного компонента, а эксперименты по устойчивости показали способность модели сохранять высокое качество в условиях шума и сдвига распределения данных.

С теоретической точки зрения, улучшение качества объясняется снижением обобщающей ошибки за счёт регуляризующего эффекта контрастивного обучения, ограничивающего сложность латентного пространства. Это обеспечивает лучшую обобщающую способность модели и её устойчивость в реальных сценариях.

Практическая значимость работы заключается в возможности применения предложенного подхода в системах рекомендаций, персонализации игрового контента и оптимизации монетизации в рамках концепции Games-as-a-Service. Полученные результаты могут быть использованы для построения интеллектуальных систем поддержки принятия решений в игровой индустрии.

В качестве направлений дальнейших исследований можно выделить расширение модели за счёт включения дополнительных модальностей, в частности эмоционального состояния пользователя, разработку методов интерпретируемости мультимодальных представлений, а также оптимизацию вычислительной сложности для масштабирования на реальные промышленные системы.

Список литературы

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
2. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9 (8), 1735-1780.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
4. Oord, A. V. D., Li, Y., & Babenko, B. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
5. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 159-170). PMLR.
6. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Dhariwal, P., Luan, M., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.
7. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2), 423-443.
8. Wang, W., Feng, Y., Chen, W., & Xu, Y. (2020). Contrastive learning for sequential recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)* (pp. 37-45).
9. Gao, J., Ge, Z., Hu, Z., & Ye, H. (2017). Deep learning for modeling user purchase

(cc) BY ©Рахимов Н.О., Дошанова М.Ю., Швец С.

behavior in mobile games. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1775-1784).

10. Wang, S., Wang, H., & Zhou, W. (2020). Multi-Modality Attention Network for User Purchase Prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (pp. 1485-1494).
11. Yang, J., Hu, M., He, Z., & Chen, J. (2022). Integrating Content and Behavior for Game Recommendation via Contrastive Learning. *arXiv preprint arXiv:2209.11737*.
12. Goldi Soni, Rohan Vaishnav, Mayank Mahant - A Comprehensive Study of the Role of Artificial Intelligence in Transforming Gaming Industry, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 13, Issue XI November 2025, DOI: https://doi.org/10.22214/ijraset.2025.74946

©Рахимов Н.О., Дошанова М.Ю., Швец С.