**УДК 004.855.5**

**Сравнение моделей и методов предиктивного кодирования в задаче восстановления мультимодальных сигналов.**

Хакимов Артём Тимурович1, Батуев Булат Базаржапович2

1 Национальный исследовательский университет «Московский физико-технический институт»,

141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9.

2 Институт радиотехники и электроники имени В.А. Котельникова Российской академии наук

125009, Россия, Москва, улица Моховая 11, корп. 7

[khakimov.at@phystech.edu](mailto:khakimov.at@phystech.edu)

**Аннотация**

За последнее десятилетие методы машинного обучения стали неотъемлемой частью подходов для обнаружения объектов, речевого интерфейса и анализа больших массивов данных. Традиционные искусственные нейронные сети (ИНС), обучаемые алгоритмом обратного распространения ошибки, демонстрируют высокую точность, однако их масштабирование сопряжено с растущими вычислительными и энергетическими затратами. Переход к более энергоэффективным архитектурам — критическое условие дальнейшего развития систем автономной обработки информации.

Импульсные нейронные сети (ИмНС) представляют собой третье поколение ИНС и предлагают принципиально иной способ кодирования: информация передаётся в виде отдельных событий-спайков, что позволяет естественным образом учитывать временную динамику сигнала и значительно сокращать число операций умножения. Энергоэффективность, асинхронная обработка и совместимость с современными нейроморфными чипами делают ИмНС привлекательными для реальных приложений. Ключевым вызовом остаётся разработка методов обучения, сочетающих биологическую правдоподобность и достаточную выразительность.

Наиболее естественным для ИмНС является семейство локальных правил пластичности, зависящих от временного распределения спайков. Классическая spike-timing-dependent plasticity (STDP) обеспечивает корректировку синаптической связи пропорционально задержке между пре- и постсинаптическими событиями, что приводит к выявлению статистически значимых корреляций во входных данных. Несмотря на привлекательность STDP как механизма обучения без учителя, для задач классификации требуется ввод внешнего сигнала ошибки либо модификация правила. Суррогатные градиентные методы частично решают задачу, однако нарушают локальность и усложняют аппаратную реализацию.

В последние годы внимание исследователей привлёк подход предиктивного кодирования, рассматривающий мозг как систему, минимизирующую локальную ошибку прогноза. В импульсной реализации этот принцип формулируется через две контр-направленные проекции: восходящую (генерация признаков) и нисходящую (предсказание входа). Локальное обновление весов по ошибке восстановления позволяет обучать сеть как с учителем, так и без него, сохраняя биологическую интерпретацию.

Наряду с алгоритмом обучения критически важен способ подачи данных. Визуальные и аудиальные сигналы можно кодировать либо пространственно, распределяя информацию по большему числу нейронов при фиксированном временном окне, либо временно, увеличивая длительность окна при ограниченном числе нейронов. Оба метода широко используются, однако их сравнительный анализ в контексте предиктивных ИмНС отсутствует.

В настоящей работе рассматриваются три взаимосвязанных аспекта. Во-первых, сравниваются пространственное и временное кодирование на единой архитектуре LIF-сети с предиктивной обратной связью. Во-вторых, оценивается информативность скрытого слоя как посредством линейной логистической регрессии, так и с помощью одномодальных и мультимодальных сверточных классификаторов. В-третьих, исследуется влияние размера скрытого слоя и постоянной времени мембраны τ на качество реконструкции и спектр спайковой активности. Предлагаемая конфигурация локальных правил (модифицированный Хеббовский закон с распадом) демонстрирует возможность получения конкурентных результатов без глобального градиента, одновременно сохраняя аппаратную реализуемость и энергоэффективность сети.

**Ключевые слова:** искусственная нейронная сеть; импульсная нейронная сеть; предиктивное кодирование; мультимодальные сигналы

**Comparison**

**models and methods of predictive coding in the problem of multimodal signal reconstruction.**

**multimodal signals.**

Khakimov Artem Timurovich, Batuyev Bulat Bazarzhapovich2

1 National Research University “Moscow Institute of Physics and Technology”,

9, Dolgoprudny, Institutskiy pereulok, Dolgoprudny, Moscow region, 141701.

2 V.A. Kotelnikov Institute of Radio Engineering and Electronics of the Russian Academy of Sciences

125009, Russia, Moscow, 11 Mokhovaya Street, Bldg. 7.

[khakimov.at@phystech.edu](mailto:khakimov.at@phystech.edu)

**Annotation**

Over the past decade, machine learning techniques have become an integral part of the algorithmic arsenal for object detection, speech interface, and big data analytics. Traditional artificial neural networks (ANNs) trained by error backpropagation algorithms have demonstrated high accuracy, but scaling them up comes with increasing computational and energy costs. The transition to more energy-efficient architectures is critical for the further development of autonomous information processing systems.

Pulse neural networks (PNNs) represent the third generation of ANNs [1] and offer a fundamentally different way of encoding: information is transmitted in the form of individual spike events, which allows us to naturally take into account the temporal dynamics of the signal and significantly reduce the number of multiplication operations. Energy efficiency, asynchronous processing, and compatibility with modern neuromorphic chips [2] make ImNSs attractive for real-world applications. The key challenge remains the development of learning methods that combine biological plausibility and sufficient expressiveness.

The family of local plasticity rules that depend on the temporal distribution of spike-timing is the most natural for ImNSs. Classical spike-timing-dependent plasticity (STDP) [6, 7] ensures that synaptic connectivity is adjusted in proportion to the delay between pre- and postsynaptic events, leading to the detection of statistically significant correlations in the input data [8]. Despite the attractiveness of STDP as a teacherless learning mechanism, classification tasks require the input of an external error signal or rule modification. Surrogate gradient methods [3-5] partially solve the problem, but violate locality and complicate hardware implementation.

In recent years, the approach of predictive coding, which considers the brain as a system minimizing the local prediction error, has attracted the attention of researchers. In an impulse realization, this principle is formulated through two counter-directional projections: bottom-up (feature generation) and top-down (input prediction). The local update of weights on the recovery error allows the network to be trained with or without a teacher, preserving the biological interpretation.

In addition to the learning algorithm, the way the data is presented is critical. Visual and auditory signals can be encoded either spatially, distributing information over a larger number of neurons with a fixed time window, or temporally, increasing the duration of the window with a limited number of neurons. Both methods are widely used, but their comparative analysis in the context of predictive ImNS is lacking.

In the present work, three related aspects are addressed. First, spatial and temporal coding on a single LIF network architecture with predictive feedback are compared. Second, the informativeness of the hidden layer is evaluated through both linear logistic regression and unimodal and multimodal convolutional classifiers. Third, the effects of the hidden layer size and membrane time constant τ on the reconstruction quality and spike activity spectrum are investigated. The proposed configuration of local rules (modified Hebbian law with decay) demonstrates the possibility of obtaining competitive results without a global gradient, while maintaining hardware realizability and energy efficiency of the network.

**Key words:** artificial neural network; pulse neural network; predictive coding; multimodal signals

1. **Введение**

За последнее десятилетие классические методы машинного обучения обеспечили существенный прогресс в задачах визуального и акустического распознавания, однако рост размеров моделей и обучающих выборок привёл к резкому увеличению вычислительных и энергетических затрат. Импульсные нейронные сети (ИмНС) — третье поколение искусственных нейронных сетей [1] — предлагают альтернативу: информация кодируется редко возникающими событиями – спайками, что уменьшает количество операций умножения и естественно отражает временную структуру сенсорного потока.

Базовым элементом ИмНС является нейрон leaky-integrate-and-fire (LIF), мембранный потенциал которого изменяется согласно

Где – постоянная времени утечки, – проводимости возбуждающих и тормозящих синапсов, – мембранный потенциал покоя, а – равновесные потенциалы синапсов. При достижении порога генерируется спайк, после чего потенциал сбрасывается до уровня потенциала покоя.

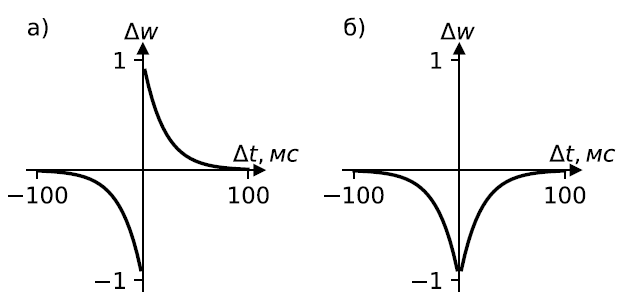
Обучение биоподобных сетей традиционно базируется на зависимой от задержки спайков пластичности (STDP). В простейшей форме изменение веса между пре- и постсинаптическим нейроном вычисляется как

что усиливает причинно-следственные связи и подавляет нерелевантные корреляции [6, 7].

В парадигме предиктивного кодирования нисходящие синапсы формируют прогноз входа, а ошибка восстановления

используется для локального обновления весов по Хеббу с распадом

Где – скорость обучения, и – активности пред- и постсинаптических нейронов соответственно, – коэффициент распада, предотвращающий неограниченный рост синаптических весов при положительной корреляции между активностями.

****

**Рис.1.** Изменение силы синаптической связи в результате обычного STDP (а) и в результате All-LTD (б)

Критически важным остаётся выбор стратегии подачи данных. Для мультимодальных сигналов возможны два предельных варианта.

1. Пространственное кодирование — увеличение числа входных нейронов при фиксированном окне (T ~ 100 мс); изображение MNIST представляется 784 нейронами, аудиосигнал — 5000 нейронами.
2. Временное кодирование — интеграция сигнала во времени при ограниченном размере входа (784 + 100 нейронов) и удлинённом окне (T ~ 1000 мс).

От выбранной схемы зависит плотность спайков, устойчивость скрытого слоя к «проклятию размерности» и точность реконструкции, измеряемая среднеквадратичной ошибкой MSE и коэффициентом корреляции. Дополнительным критерием является информативность латентного представления, оцениваемая логистической регрессией и сверточными классификаторами (1D-CNN и мультимодальная CNN).

Целью данной работы является провести систематическое сравнение пространственного и временного кодирования в предиктивной ИмНС при восстановлении изображений и аудио, а также количественно оценить репрезентативность скрытого слоя различными классификаторами.

1. **Экспериментальная часть**

* 1. *Архитектура модели*

Предлагаемая импульсная сеть состоит из трёх функциональных популяций: двух сенсорных входов (визуального и аудиального) и общего скрытого слоя. Визуальный канал представлен ансамблем из 784 LIF-нейронов, каждому из которых соответствует пиксель изображения MNIST; аудиальный канал формируется либо 5000, либо 100 нейронами в зависимости от выбранной схемы кодирования. Оба входных ансамбля проецируются в скрытый слой через прямые возбуждающие синапсы. Потенциал каждого скрытого нейрона определяется суммой взвешенных импульсов, поступающих от изображений и аудио, что обеспечивает раннюю интеграцию модальностей в едином латентном пространстве.

От скрытого слоя к каждому входному нейрону проведён симметричный набор «обратных» синапсов. Эти нисходящие связи формируют прогноз сенсорного сигнала и, будучи аддитивно сложены с реальным входом, позволяют вычислить локальную ошибку восстановления. Таким образом реализуется классическая концепция предиктивного кодирования: вверх по иерархии передаются лишь непрогнозируемые компоненты стимула, а вниз — ожидание ближайшего сенсорного состояния.

Синаптическая пластичность задаётся модифицированным правилом Хебба с распадом. Вес увеличивается, когда пресинаптический и постсинаптический нейроны активны синхронно, и экспоненциально уменьшается пропорционально собственной величине, что предотвращает неограниченный рост и способствует стабилизации обучения. Одно и то же правило применяется к прямым и обратным проекциям, поэтому модель обучается локально, без глобального распространения ошибки, и сохраняет биологическую правдоподобность.

Изображение выглядит как текст, диаграмма, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

**Рис.2.** Схема мультисенсорного предиктивного кодирования. Мультисенсорность реализована за счет интегрирования аудио и видео сигналов в скрытый слой.

Такой подход обеспечивает согласованную обработку мультимодальных потоков, устойчивость к различным режимам кодирования (пространственному и временному) и энергоэффективность, характерную для импульсных нейросетей.

* 1. *Методика экспериментов*

Для воспроизведения результатов эксперимента существенно фиксировать все численные настройки сети и процесса симуляции. Ниже приведены ключевые гиперпараметры для двух режимов подачи данных, определяющие конфигурацию входных слоёв, скрытого слоя и динамику обучения. Условия проведения эксперимента:

Датасеты:

* MNIST: 60 000 обучающих и 10 000 тестовых изображений 28×28.
* Free Spoken Digit Dataset (FSDD): 2000 аудиофайлов цифр 0–9; из них сформировано 1000 синхронизированных пар «изображение + аудио».

Предварительная обработка:

Изображения:

* Деление пикселей на 255 → масштабирование к [0,1].
* Дополнительное деление на 4 — ограничение максимальной частоты спайков.

Аудио:

* Удаление тишины и нормировка амплитуды к [0,1].
* Усечение либо дополнение до 5000 отсчётов.

Ключевые параметры модели:

Пространственное кодирование

* вход: 784 нейрона (изображение) + 5000 нейронов (аудио)
* скрытый слой: 370 нейронов
* окно симуляции: 100 мс
* постоянная времени: τ = 10 мс
* скорость обучения: η = 10⁻⁶ Hz
* распад: λ = 4·10⁻³

Временное кодирование:

* вход: 784 нейрона (изображение) + 100 нейронов (аудио)
* скрытый слой: 400 нейронов
* окно симуляции: 1 c
* постоянная времени: τ = 20 мс

1. **Результаты и обсуждения**

Точность реконструкции оценивалась по двум классическим критериям: среднеквадратичной ошибке и коэффициенту линейной корреляции. Для визуальных данных каждой цифры исходное изображение нормировали к диапазону [0, 1] и сравнивали с восстановленным кадром покадрово; MSE вычисляли как усреднённую по 784 пикселям разность квадратов, после чего усредняли значения по всем тестовым примерам. Аналогичная процедура применялась к аудиофрагментам: нормированный временной ряд длиной 5000 отсчётов сопоставляли с реконструкцией по отсчётам, а результирующее MSE усредняли по 100 невиданных записям. Дополнительно, чтобы зафиксировать сохранение относительных амплитудных паттернов, между исходным и восстановленным сигналами рассчитывали корреляцию Пирсона; для изображений это делалось по векторизованному массиву яркостей, для аудио — по всему временно́му окну. Полученные значения 0.02–0.04 MSE и 0.96–0.98 корреляции для изображений, а также 0.07–0.10 MSE и 0.89–0.94 корреляции для аудио показывают, что модель надёжно сохраняет как геометрию цифр, так и основную амплитудно-частотную структуру речевого сигнала.

Информативность латентного пространства проверяли тремя независимыми методами. Сначала для каждого примера усредняли потенциалы скрытых нейронов по времени, формируя вектор длины на этой выборке обучали логистическую регрессию и фиксировали её accuracy на отложенных данных (0.74–0.79). Затем использовали одномерную сверточную сеть: в качестве признаков подавали полный временной профиль потенциалов, что позволило задействовать динамическую информацию и незначительно повысить точность до 0.76. В качестве контрольного ориентира обучали мультимодальную CNN, получающую на вход исходные изображения и аудио без предиктивного слоя; её результат (0.46) оказался существенно ниже, что подтверждает: скрытое представление, сформированное предиктивной спайковой сетью, содержит существенно более компактное и устойчивое описание стимула, нежели прямое сверточное извлечение признаков при аналогичном объёме обучающих данных.

**Изображение выглядит как пиксель, прямоугольный

Контент, сгенерированный ИИ, может содержать ошибки.**

Изображение выглядит как снимок экрана, График, текст, линия

Контент, сгенерированный ИИ, может содержать ошибки.**Рис.3.** Оригинал изображения и реконструированное (восстановленное) изображение для цифры 3 **Рис.4.** Исходный аудиоряд и его реконструкция (восстановление) для цифры 3. Шум отчетливо видно начиная с 2000 отсчета.

Отдельного внимания заслуживает характер шума, проявляющийся в реконструированном аудиосигнале (см. рис. 4). После активной голосовой фазы модель воспроизводит затухающую квазистохастическую компоненту, амплитуда которой остаётся заметной вплоть до конца окна, тогда как оригинальный сигнал быстро переходит в тишину. Наиболее правдоподобное объяснение связано с «перекодировкой избыточности»: аудиовход описывается 5000 независимыми LIF-нейронами, поэтому при усреднении по времени каждый нейрон получает очень малую долю информативных спайков. Часть входных нейронов практически не активируется, другая — генерирует редкие одиночные спайки, которых оказывается достаточно для того, чтобы обратная проекция скрытого слоя поддерживала слабый, но постоянный прогноз. Фактически модель воспринимает собственный шум как предсказуемую составляющую, а затем пытается «объяснить» его через латентные переменные, что и проявляется в остаточном шуме.

Для проверки того, насколько скрытый слой предиктивной спайковой сети действительно концентрирует семантическую информацию о классе стимула, были применены три независимых способа пост-оценки.

Во-первых, потенциалы всех скрытых нейронов усреднялись по времени внутри каждого предъявления. Полученный вектор размерности служил признаковым описанием объекта; на этой компактной выборке обучалась логистическая регрессия с L2-регуляризацией. Такой линейный классификатор продемонстрировал точность 75 % на тестовой подвыборке, что заметно превосходит случайный уровень и подтверждает наличие читаемой, пусть и не идеально линейной структуры в латентном пространстве.

Во-вторых, чтобы установить верхнюю границу точности для «сырых» данных, аналогичная логистическая регрессия была обучена непосредственно на нормированных пикселях изображений MNIST (без использования аудио). В этом случае точность классификации достигла 91 %, поскольку изображение содержит полную статическую информацию о классе, тогда как усреднённый скрытый слой представляет лишь сжатую версию мультимодального входа. Разница в 16 % количественно показывает, какая доля информации теряется при переходе от полного сенсорного паттерна к спайк-кодированному представлению.

Наконец, для проверки того, не обусловлено ли снижение точности самой мультимодальной природой входа, была построена сверточная эталонная модель. Мультимодальная CNN включает две параллельные ветви: 2D-свёртки для изображения и 1D-свёртки для аудио, после чего признаки объединяются суммированием и передаются в полносвязный слой размерности 300. Несмотря на большее число параметров и прямой доступ к исходным данным, такая модель достигла лишь 46 % точности. Это свидетельствует о том, что при ограниченном объёме обучающих примеров прямое сверточное объединение модальностей склонно к переобучению и не извлекает устойчивых признаков. Предиктивная же SNN, напротив, благодаря биологически мотивированному механизму локальной ошибки восстановления, формирует более выразительное и компактное скрытое представление.

Таким образом, совокупность результатов логистической регрессии и мультимодального сверточного контроля показывает следующее: скрытый слой действительно содержит информацию о классе стимула, часть дискриминативных признаков теряется при преобразовании сигнала, но даже в сжатом виде латентное представление, сформированное предиктивной моделью, оказывается надёжнее и устойчивее, чем признаки, извлечённые классическими CNN.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Контент, сгенерированный ИИ, может содержать ошибки.

**Рис.5.** Схема оценки информативности скрытого слоя модели. Слева: латентное представление, сформированное в спайковой SNN, усредняется по времени и используется для обучения логистической регрессии. Справа: эталонная сверточная архитектура, принимающая изображение и аудио параллельно через Conv2D и Conv1D-ветви, объединяющиеся на полносвязном уровне.

Сравнение динамики обучения классического и спайкового автоэнкодеров демонстрирует качественно различное поведение метрики ошибки. На втором графике (классическая модель) видно плавное, экспоненциально затухающее снижение среднеквадратичной ошибки (MSE) от значения выше 0.06 до уровня ниже 0.01 уже за первые 100 эпох. График является гладким, с устойчивой динамикой сходимости и практически монотонным улучшением качества. Это соответствует типичному поведению MLP-автоэнкодера, обучаемого на непрерывных входах с использованием градиентных методов и функции активации типа сигмоида.

На первом графике (спайковая реализация) наблюдается иная картина: снижение ошибки идёт медленно, с выраженными плато, ступенчатыми сдвигами и периодами стагнации. Даже к 50-й эпохе ошибка остаётся на уровне ~ 0.0304–0.0307, что существенно выше финального значения у классического автоэнкодера. Такая форма кривой сходимости объясняется несколькими факторами. Во-первых, в спайковом автоэнкодере обучение происходит с использованием локальных правил пластичности, основанных на синаптической коактивации, без прямого доступа к градиенту функции потерь. Во-вторых, информация подаётся в дискретной, событийной форме (через спайки), и потенциалы скрытого слоя обновляются с временным лагом, что снижает чувствительность к мелким ошибкам. Кроме того, архитектура использует leaky-integrate-and-fire нейроны, чья динамика ограничивает разрешающую способность по амплитуде и временным частям сигнала.

Изображение выглядит как линия, График, диаграмма, число

Контент, сгенерированный ИИ, может содержать ошибки.

**Рис.6.** Сравнение динамики обучения классического (слева) и спайкового (справа) автоэнкодеров на задаче реконструкции изображений MNIST. На графиках представлена зависимость среднеквадратичной ошибки (MSE) от номера эпохи.

1. **Выводы**

В работе реализована и проанализирована модель импульсной нейронной сети, реализующая механизм предиктивного кодирования для восстановления мультимодальных сигналов. Показано, что при обучении на базе локального правила Хебба с распадом сеть успешно восстанавливает как визуальные, так и аудиальные входы, демонстрируя на изображениях MNIST низкий уровень ошибки (MSE ≈ 0.03) и высокую корреляцию (r > 0.96). Пространственное кодирование оказалось более устойчивым и точным по сравнению с временным, особенно на аудиосигналах, где высокое число входных нейронов приводит к эффекту «проклятия размерности». Скрытый слой формирует латентное представление, пригодное для классификации: логистическая регрессия по его активности достигает точности 75 %, а сверточная 1D-CNN подтверждает наличие временной структуры. По сравнению с мультимодальной CNN, обучаемой end-to-end, предиктивная SNN демонстрирует более устойчивое и компактное кодирование при сопоставимых или лучших результатах. Работа подтверждает применимость локального обучения и событийной обработки для задач интеграции и восстановления сенсорных мультимодальных сигналов.

**Литература**

1. Antsiperov V. E. Generative Model of Autoencoders Self-Learning on Images Represented by Count Samples // Automation and Remote Control. 2022. Vol. 83. No. 12. Pp. 1959–1983. DOI: 10.1134/s00051179220120098.
2. Abbott L. F. Lapicque’s introduction of the integrate-and-fire model neuron (1907) // Brain Research Bulletin. 1999. Vol. 50. No. 5–6. Pp. 303–304. DOI: 10.1016/s0361-9230(99)00161-6.
3. Friston K. The free-energy principle: A unified brain theory? // Nature Reviews Neuroscience. 2010. Vol. 11. No. 2. Pp. 127–138.
4. LeCun Y., Cortes C., Burgess C. J. C. MNIST handwritten digit database // AT&T Labs. 2010. [Online]. URL: <http://yann.lecun.com/exdb/mnist>
5. Trappenberg T. P. Fundamentals of Computational Neuroscience. 3rd ed. Oxford University Press, 2022.
6. Zohar J. Free Spoken Digit Dataset (FSDD) [Electronic resource]. 2020. URL: <https://github.com/Jakobovski/free-spoken-digit-dataset>
7. Predictive Coding Multisensor v3 [Electronic resource]. URL: <https://github.com/buligar/Predictive-coding/blob/main/Predictive%20coding%20Multisensor%20v3.ipynb>
8. Gerstner W., Kistler W. M. Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press, 2002.
9. Zenke F., Ganguli S. Superspike: Supervised learning in multilayer spiking neural networks // Neural Computation. 2018. Vol. 30. No. 6. Pp. 1514–1541. DOI: 10.1162/neco\_a\_01086.
10. Brea J., Urbanczik R., Senn W. Prospective coding by spiking neurons // PLoS Computational Biology. 2016. Vol. 12. No. 6. e1005003. DOI: 10.1371/journal.pcbi.1005003.
11. Tavanaei A., Ghodrati M., Kheradpisheh S. R., Masquelier T., Maida A. Deep learning in spiking neural networks // Neural Networks. 2019. Vol. 111. Pp. 47–63. DOI: 10.1016/j.neunet.2018.12.002.
12. Brette R. Philosophy of the spike: Rate-based vs. spike-based theories of the brain // Frontiers in Systems Neuroscience. 2015. Vol. 9. P. 151. DOI: 10.3389/fnsys.2015.00151.