

# Эволюция нейроморфного ИИ

Возможна ли альтернатива LLM

Андрей Лаврентьев

Руководитель отдела развития технологий

# Нейроморфный ИИ. Основные принципы

- Вычисления в памяти
- Локальность
- Импульсные сети

## LLM. Основные принципы

- Токенизация
- Embedding
- Positional Encoding
- Attention (Q, K, V)
- Многослойность
- Генеративность

## LLM. Ресурсоемкость

- $d$  – большая размерность эмбеддинга и позиционного кодирования
- $L$  – длинный контекст для удержания внимания
- $L^2$  – квадратичная сложность вычисления Attention
- Backprop – обучение через глобальный или LoRA – через частичных проход по слоям
- Прогноз по одному токену – вероятностный характер, легко сбить, приводит к галлюцинациям
- Все слои ~ одинаково ресурсоемки

## LLM. Исследования

- Дистилляция
- Выделение «важных» весов в матрицах внимания и в полносвязанных слоях
- Переход на низко битные веса, или даже на бинарные
- Использование расширяющих автоэнкодеров для анализа работы сети
- Поиск альтернатив трансформерам для обхода квадратичной сложности (KV Cache, MoE, SSM – Mamba/Jamba, Window attn, ...)

## LLM. Предельный переход

- Событийные связи между слоями -> импульсы
- Ассоциативный характер таких связей
- Внимание – как выбор лучшего из множества локальных для слоя вариантов
- Прозрачная интерпретация без расширяющих автоэнкодеров
- Распараллеливание
- Бесконечный контекст

# Нейроморфный ИИ. Технологии

- Временная и пространственная суммация  $\sum_t \sum_i spikes$
- Масштабируемые колончатые архитектуры ColaNet
- Локальные вычисления  
на уровне нейрона - STDP и на уровне популяции - WTA
- Механизм пространственной локализации на уровне популяции для синхронизации на интервале времени - Astrocytes
- Поддержка вычислений на уровне нейроморфных чипов  
ориентированных на событийные вычисления в памяти AltAI

# Технологии сегодня на стендах и в докладах

- Kaspersky
- СВ – обучение
- МИО-сигналы
- Айтрэкер
- ВСІ-интерфейс
- Астроциты
- Spike LLM
- Отраслевые
- Event Processor (EP)

The collage includes:

- A screenshot of the Kaspersky Machine Learning for Anomaly Detection interface.
- A screenshot of the SpikeLLM web interface showing a chat about information security.
- A diagram illustrating the RAG (Retrieval-Augmented Generation) process.
- A diagram of a neural network architecture labeled "Spike" with a bar chart overlay.
- A screenshot of a presentation slide titled "We show that astrocyte can act as...".
- A video frame showing a person in a lab coat.
- A screenshot of a game or simulation interface.

# Нейроморфный ИИ. Нейросемантическая сеть в EP

- Пространственно-временной векторный вход  $\Sigma_{t,i} spikes$
- Нейрогенез – идентификация, новизна, постоянная память
- Обучение за счет минимизации ресурсов сети – компактификация хранимых пространственно-временных паттернов
- Внимание – как внешне задаваемое, так и внутреннее
- Слой для «слушателей» сети - внешнее внимание
- Рекурсивные сети - внутреннее внимание – синхронизация
- Семантика – как свойство синхронизации на разных временных интервалах на разных слоях сети

# Вход нейросемантической сети



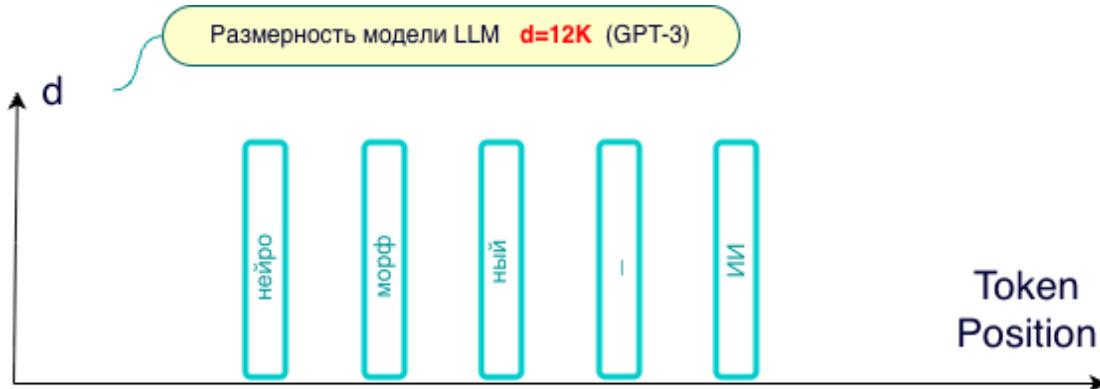
Вход нейрона  $N_l^i$  на слое  $l$  - пространственно-временной вектор из идентификаторов нейронов предыдущего слоя  $N_{l-1}^i$  и интервалов  $d_i = t_i - t_{i-1}$  относительно предыдущих нейронов :

$$[ (N_{l-1}^1, 0), (N_{l-1}^2, t_2 - t_1), \dots, (N_{l-1}^m, t_m - t_{m-1}) ]$$

Если данные = текст, то для нулевого слоя нейроны  $N_0^i$  соответствуют одному символу с условной единичной длительностью

$$t_i = t_{i-1} + 1$$

# Вход трансформерного блока LLM



Входной контекст разбит на токены: "нейро" "морф" "ный" "\_" "ИИ"  
 Каждый токен преобразован в эмбеддинг-вектор размерности  $d$ .  
 Каждая позиция преобразована в вектор той же размерности  $d$ .  
 Итоговый сэмпл на входе - сумма векторов эбединга и позиции.

Позиции токенов задают порядок.

Порядок отражен в векторном представлении позиции каждого токена (PE, RoPE)

Если нужно не только порядок, но и длительность интервалов между позициями  $\Delta t$ , то ее вводить нужно отдельным свойством (RoPE + $\log(\Delta t)$ , RoMAE)

# Сложность

	LLM (GPT-3 175 B)	Нейросемантическая сеть		
<b>Контекст</b>	В токенах $N_{vocab} = 50k$ Фиксированная длина	$L \sim 2k$ $L \sim 128k$ (2025 г)	В символах $N_{vocab} \sim 100$ открытый словарь Переменная длина - эпизод	$L \leq 4k$
<b>Цель</b>	Охватить вниманием как можно больший контекст		Обработать эпизод в привязке к предыдущим эпизодам с учетом внимания	
<b>Элемент на входе 0-слоя</b>	Вектор размерности $d = Emb + Pos Enc$	$d \sim 12k$	Единичный пространственно-временной вектор	$k = 1$
<b>Элемент на входе след. слоев</b>	Вектор размерности:	$d \sim 12k$	Пространственно-временной вектор размерности:	$k = [1, 8]$
<b>Вход слоя</b>	Тензор	$L * d$	Пространственно-временной вектор из идентификаторов нейронов предыдущего слоя $N_{l-1}^i$ и их времен $t_i$ : $[(N_{l-1}^1, t_1), \dots, (N_{l-1}^m, t_m)]$	$D_{l-1}$ – количество элементов в контексте предыдущего слоя $D_0 = L$ $D_l \sim D_{l-1} / k$
<b>К-во слоев</b>	Слой = трансформер блок	96	Слой = нейроны + WTA + до 5 рекурсивных сетей внимания	$48 = 8 + 5 * 8$
<b>К-во нейронов на слое</b>	$const \sim 6:$ $1_Q + 1_K + 1_V + 1_{Prj} + 1_{MLP} + 1_{BN}$	$\sim L * const$	Нейрогенез. Насыщение определяется предметной областью	$\hat{D}_l$ - уникальных элементов по всем контекстам слова
<b>Сложность на слое</b>	$K * Q$	$L^2$	Свертки по входу + WTA по выходу	$kD_{l-1} + \ln(kD_l)$
<b>Операции</b>	Перемножение матриц размерности	$\sim L * d$	Разреженные вычисления на пространственно-временных векторах и сужающемся контексте	$\sim k * D_l$

# Event Processor - продукт на основе NSN



Раннее обнаружение  
атак



Анализ на потоке



Детектирующие сценарии  
без четких критериев



Профилирование  
трафика

# Компоненты и данные

## Системы объекта мониторинга



# Элементы нейросемантической сети



Нейроны

Обеспечивают идентификацию событий и паттернов, хранение истории, точное воспроизведение



Головы  
внимания

Фильтруют поток событий согласно своему вниманию и выучивают паттерны поведения



Мониторы

Реализуют сценарии мониторинга на основе событий, паттернов и статистики, шлют алерты.

# Пример 1

Детектирование входа с необычного хоста



Классические правила корреляции не помогут



Используем базовые возможности Event Processor:

Идентификация события

Новизна значения поля «адрес источника» в событии

Внимание на выбранных или на всех пользователей

## Пример 2

### Активность в группе в необычное время



Пересчитывать статистику раз в день – опоздаем с детектом



Используем базовые и статистические возможности Event Processor:

Attention Head выделяет группу пользователей  
Stat-Monitor сравнивает каждое событие со статистикой, и сразу  
обновляет статистику

# Пример 3

## DNS эксфильтрация



Многопараметрическая задача не решается правилами



Используем Stat-Monitor с признаками:

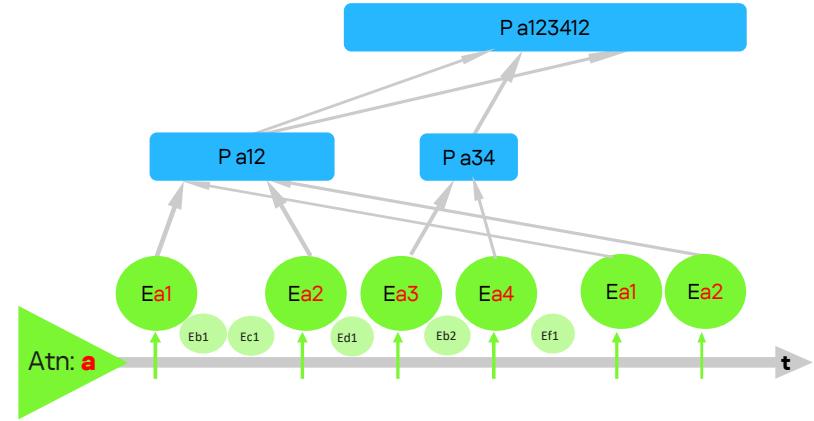
- хост
- домен
- длина
- энтропия
- частота каждого символа

Статистика обновляется с каждым событием

# Нейросемантическая сеть. Преимущества

Q

- ✓ Учится постоянно
- ✓ Понимает «новизну»
- ✓ Имеет постоянную память
- ✓ Не галлюцинирует
- ✓ Дает прозрачную интерпретацию



# Нейросемантическая сеть. Аппаратные потребности

- ✓ Нейрогенез → динамическое изменение архитектуры
- ✓ Постоянная память -> много нейронов ( $10^6 - 10^9$ )
- ✓ Много нейроморфных ядер (ядро ~ 512 нейронов)
- ✓ WTA – локальные вычисления задействуют несколько ядер
- ✓ Внешнее внимание и рекурсивные сети –  
интенсивная маршрутизация сообщений
- ✓ Энергопотребление = 2 милливатта на ядро  
 $10^6$  нейронов ->  $2 \cdot 10^3$  ядер -> 4 Вт  
 $10^9$  нейронов ->  $2 \cdot 10^6$  ядер -> 4 кВт