Санкт-Петербургский государственный университет

Применение дискретизованных нейросетей

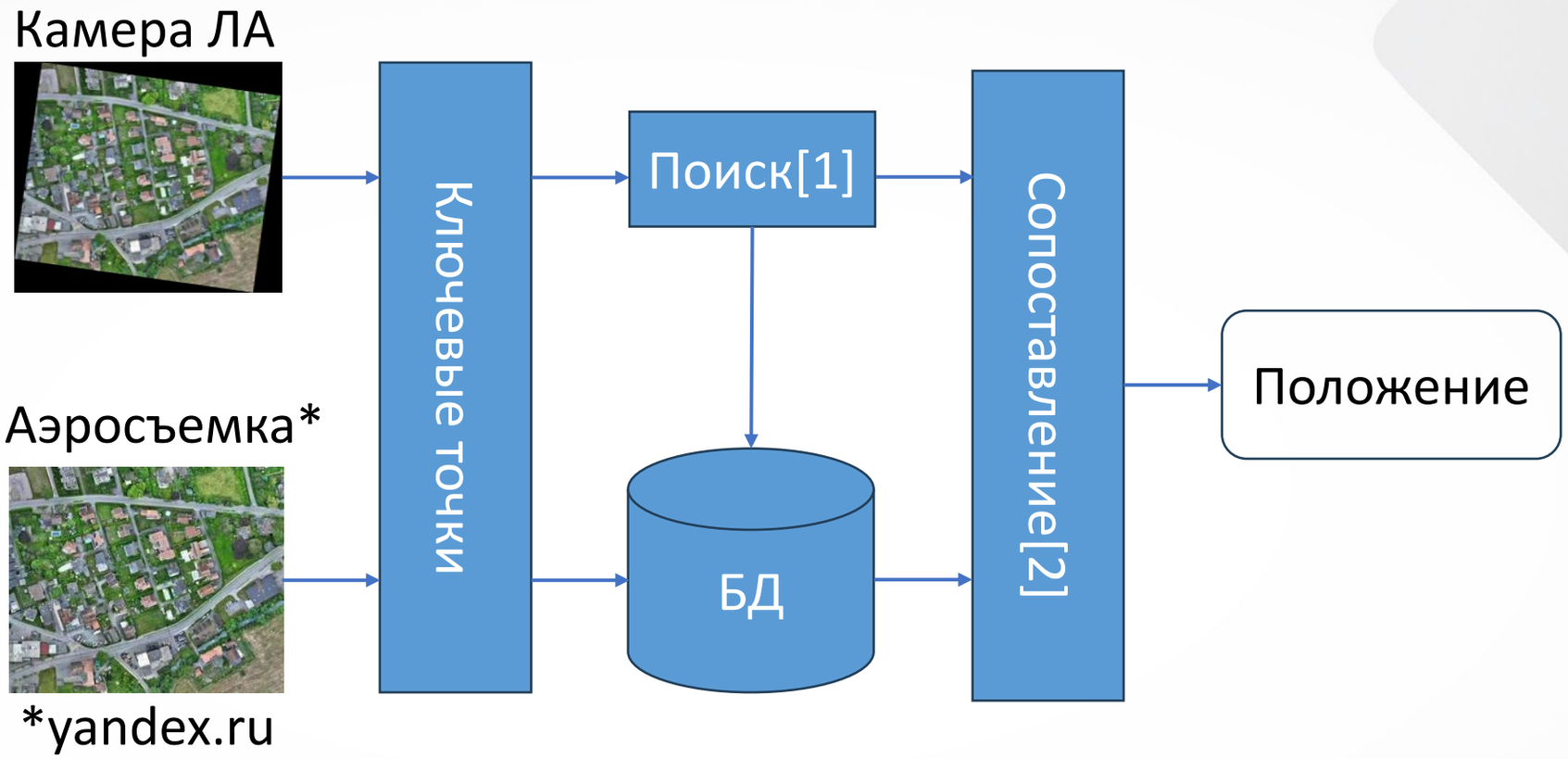
к. ф.-м. н. ст. преп. каф. информатики СПбГУ
Салищев Сергей Игоревич

Основные тезисы

- Нейросеть эффективно вычислима
  - Нейрочипы
- Нейросеть легко интерпретируема
  - Ансамбль деревьев решений
- Нейросеть – дискретный объект
  - Шум квантования
  - Информационная емкость
- Нейросеть эффективно обучается
  - Пакетный стохастический градиентный спуск
  - Остаточные сети и пакетная нормализация

Санкт-Петербургский государственный университет

Пример: Система бесспутниковой навигации



[1] Yang Y., Newsam S. Geographic image retrieval using local invariant features //IEEE transactions on geoscience and remote sensing. – 2012. – Т. 51. – №. 2. – С. 818-832.
[2] Xiong Z., Zhang Y. A critical review of image registration methods //International Journal of Image and Data Fusion. – 2010. – Т. 1. – №. 2. – С. 137-158.

Добавляем нейросеть

- Основные недостатки системы на основе классических алгоритмов машинного зрения
  - Высокая вычислительная сложность сопоставления из-за низкого качества признаков
  - Низкая точность/полнота поиска
  - Расчеты на CPU
- Заменим на нейросеть
  - Поиск ключевых точек и вычисление признаков [1]
  - Сопоставление (не обязательно) [2]

[1] Song W., Li S., Benediktsson J. A. Deep hashing learning for visual and semantic retrieval of remote sensing images //IEEE Transactions on Geoscience and Remote Sensing. – 2020. – Т. 59. – №. 11. – С. 9661-9672.

[2] Nassar A. et al. A deep CNN-based framework for enhanced aerial imagery registration with applications to UAV geolocalization //Proceedings of the IEEE conference on computer vision and pattern recognition workshops. – 2018. – С. 1513-1523.

Санкт-Петербургский государственный университет

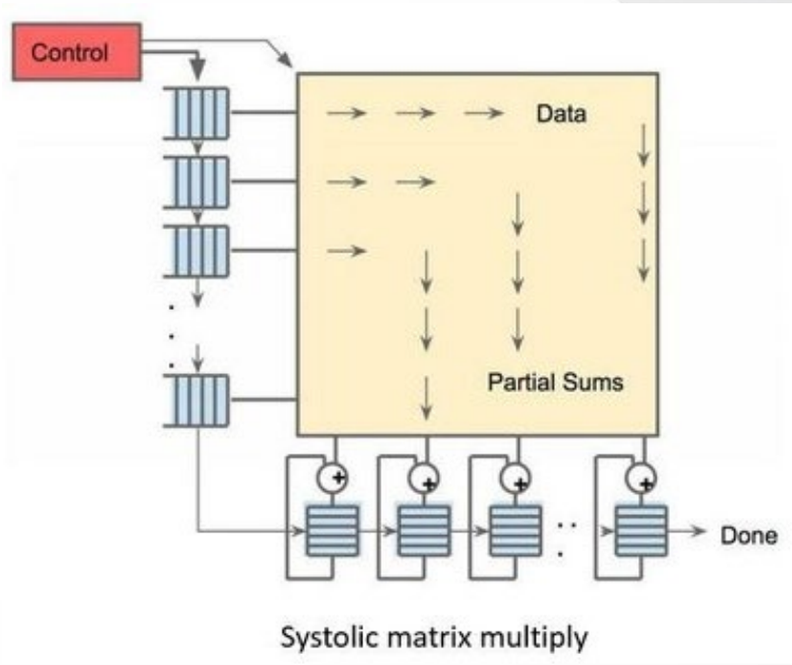
Пример встроенной вычислительнойплатформы

- Rockchip https://www.rock-chips.com RK3566/RK3568, RK3588, RK3562
  - 4-8 ядер ARM Cortex-A 1-2 ГГц
  - GPU ARM
  - Типичное энергопотребление 5Вт, макс. 20Вт
- Примерный вычислительный бюджет
  - CPU – 4 GFLOPS $= 4 \cdot 10^9$оп/с, FP32 (Процессор)
  - GPU – 40 GFLOPS $= 4 \cdot 10^{10}$оп/с, FP32 (Графический ускоритель)
  - TPU – 4 TOPS $= 4 \cdot 10^{12}$ оп/сек, 8 бит MAD (Нейрочип)
- Практическое определение Нейросети?
  - Все что вычисляется на нейрочипе

Санкт-Петербургский государственный университет

Почему TPU в 1000 раз быстрее CPU

- Пониженная точность вычислений
- Однородные матричные вычисления
- Кэширование данных на чипе
- Сверхшироки интерфейс локальной памяти



Systolic matrix multiply

Санкт-Петербургский государственный университет

Средства разработки Rockchip

- Обучение нейросети
  - PyTorch, TensorFlow, etc
- Промежуточный формат ONNX
- Прикладной интерфейс RKNN-Toolkit2 https://github.com/airockchip/rknn-toolkit2
  - Языки встроенной платформы C/C++, Python
  - ОС Linux, Android
- Отладка и симуляция на ПК RKNPU2 https://github.com/airockchip/rknn-toolkit2/tree/master/rknpu2

Санкт-Петербургский государственный университет

Так что же такое нейросеть

- Непрерывный и гладкий объект

$$x_{i+1} = f(W_i x_i + b_i)$$
$$x_i \in R^{n_i}, f \in \mathrm{C}^1$$

- Матрица W имеет разреженную блочную периодическую структуру
- Обучается градиентным спуском
- f – произвольная гладкая функция кроме степенной

Пусть $L(\theta) \to min$ - целевая функция, γ – скорость обучения

$$g_t = \nabla_\theta L(\theta_{t-1})$$
$$\theta_t = \theta_{t-1} - \gamma_t g_t$$

Санкт-Петербургский государственный университет

Свойства нейросетей

- Нейросеть - универсальный ε аппроксиматор для любых интегрируемых по Лебегу функций нескольких переменных, нелинейность любая кроме полинома
  - Достаточно 3 слоев (ширина не ограничена) [1]
  - Достаточно $n + 1$ каналов (глубина не ограничена) [2]

[1] Ismailov V. E. A three layer neural network can represent any multivariate function //Journal of Mathematical Analysis and Applications. – 2023. – Т. 523. – №. 1. – С. 127096.

[2] Kratsios A., Papon L. Universal approximation theorems for differentiable geometric deep learning //The Journal of Machine Learning Research. – 2022. – Т. 23. – №. 1. – С. 8896-8968.

Санкт-Петербургский государственный университет

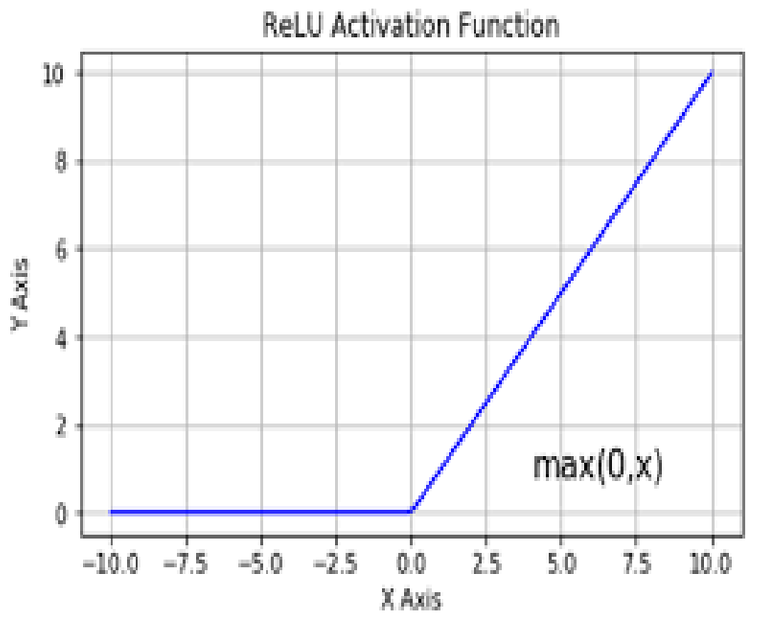
Мотивация глубоких нейросетей

- Уменьшение размерности внутренних слоев за счет факторизации матриц W
  - Обучение глубоких нейросетей неустойчиво из-за исчезновения градиента
- Качественно обученная глубокая нейросеть – эффективный метод ускорения вычислений для любых алгоритмов!

Санкт-Петербургский государственный университет

Убираем гладкость

- Упрощаем нелинейность
- $f(x) = ReLU(x) = \max(0, x)$
- Сеть с ReLU эквивалентна дереву решений[1]

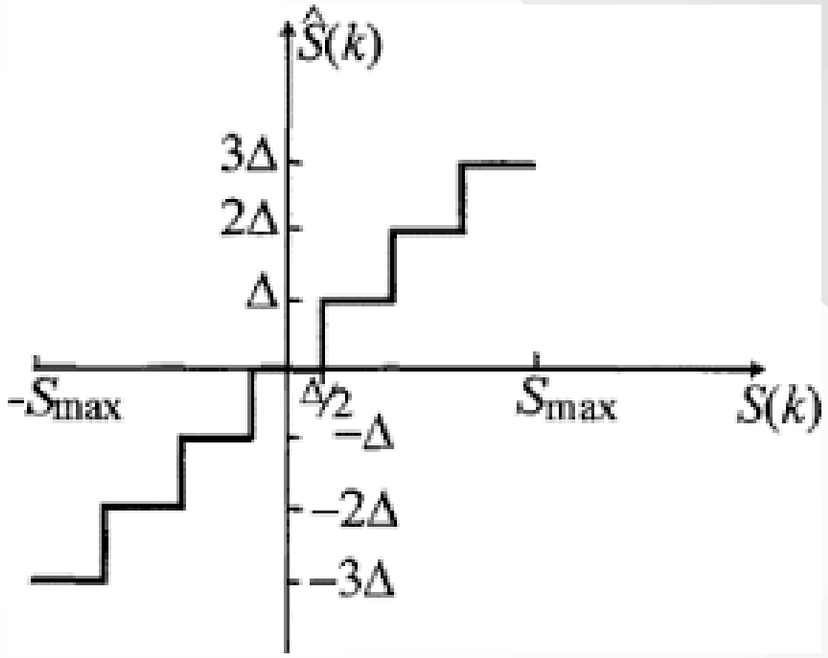


[1] Aytekin C. Neural Networks are Decision Trees //arXiv preprint arXiv:2210.05189. – 2022.

Санкт-Петербургский государственный университет

Убираем непрерывность

- $x_{i+1} = f\left((W_i + u_i)x_i + b_i\right) + v_i$
  - u_i, v_i – шум
- Вычисления с фиксированной точкой малой разрядности
  - Шум – погрешность округления
  - Сокращение длины арифметического переноса между разрядами
- Спайковые сети
  - Моделирование биологических нейросетей
- Вычисления в памяти
  - Нейросети на чипе
  - Физическая нечеткая логика



Санкт-Петербургский государственный университет

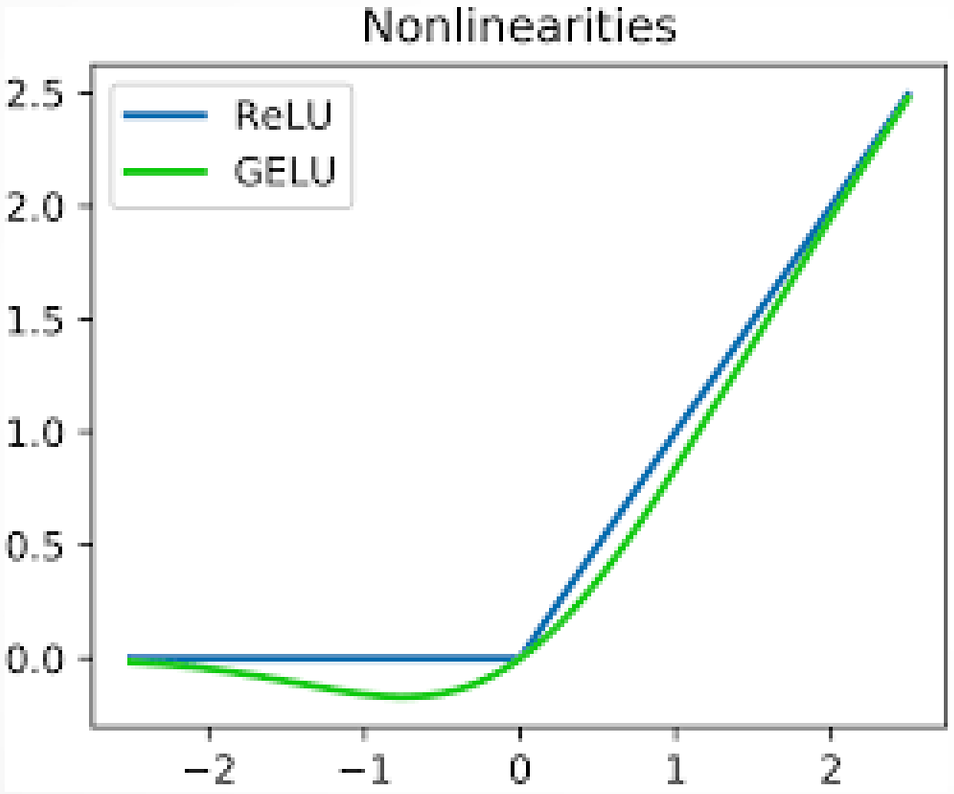
Градиентный спуск? Контрастное обучение

- В биологических нейросетях градиентный спуск (пока) не найден[1]
- Как тренировать нейросети без ГС[1]?
  - Послойное контрастное обучение
  - Мешок признаков
    - Сначала генерируем признаки, потом вешаем метки
    - Вероятность линейного разделения кластеров → 1 с ростом числа признаков
    - Как научиться попадать в мишень
      1. Научится попадать кучно
      2. Сдвинуть прицел на центр мишени
- Градиентный спуск (в разы) быстрее
- Обучение нейросетей в природе интерпретируется как контрастное обучение

[1] Hinton G. The forward-forward algorithm: Some preliminary investigations //arXiv preprint arXiv:2212.13345. – 2022.

Санкт-Петербургский государственный университет

Регуляризация, в поисках потерянной гладкости



Санкт-Петербургский государственный университет

Пакетный Стохастический градиент

$$L(w) = \frac{1}{n}\sum_i L\left(w, x_{(i)}\right)$$

$$\nabla L(w) = \frac{1}{n}\sum_i \nabla L\left(w, x_{(i)}\right) \to E\nabla L\left(w, x_{(i)}\right), n \to \infty$$

Основная мотивация пакетного (batch) градиента – распараллеливание обучения для более эффективных вычислений

Требуется непрерывность $E\nabla L\left(w, x_{(i)}\right)$ а не $\nabla L(w)$

Устойчивость к малому шуму в данных → Непрерывность и Ограниченная вариация EL?
Ограниченная вариация EL → Гладкость EL почти везде

Санкт-Петербургский государственный университет

Основные причины неустойчивого обучения

- Спящие нейроны

$$\left(x_{i,k}\right)_t = Const$$

- Пропадающий градиент в результате декорреляции

$$x_t \sim RV$$

Санкт-Петербургский государственный университет

Улучшение устойчивости обучения

- Групповая нормализация (batch norm) I

$$b_{i,k} = -\mathrm{E} W_{i,k} x_i$$

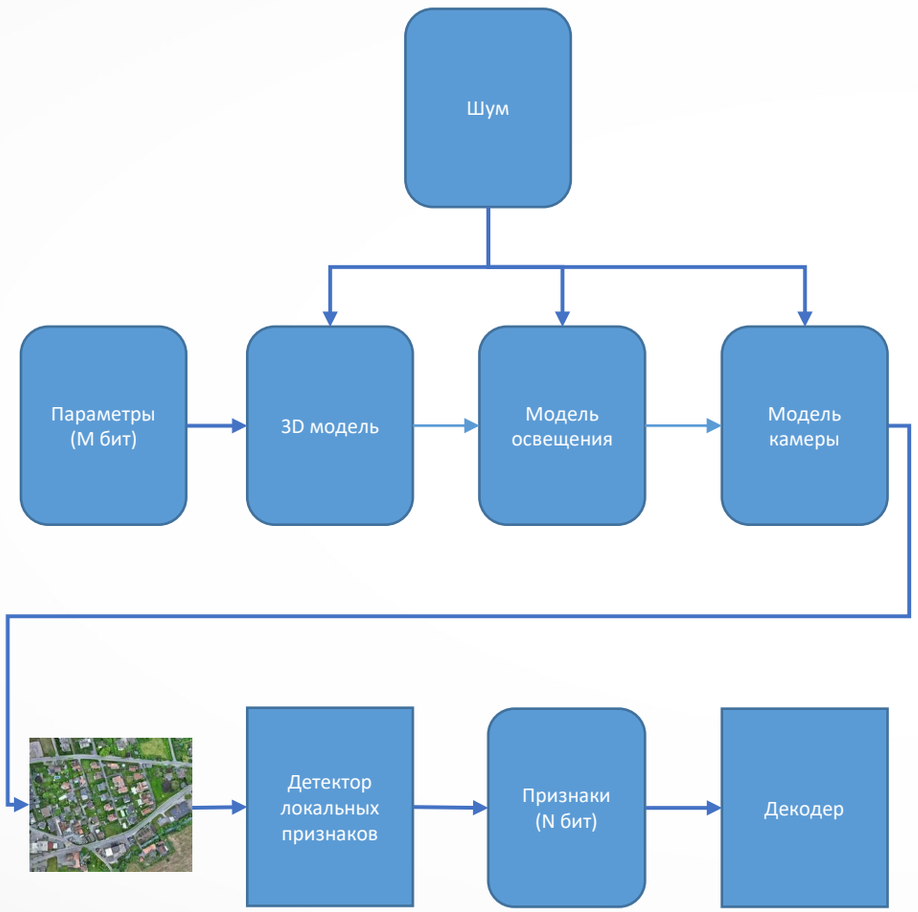
- Остаточные сети (residual network, skip connection) II

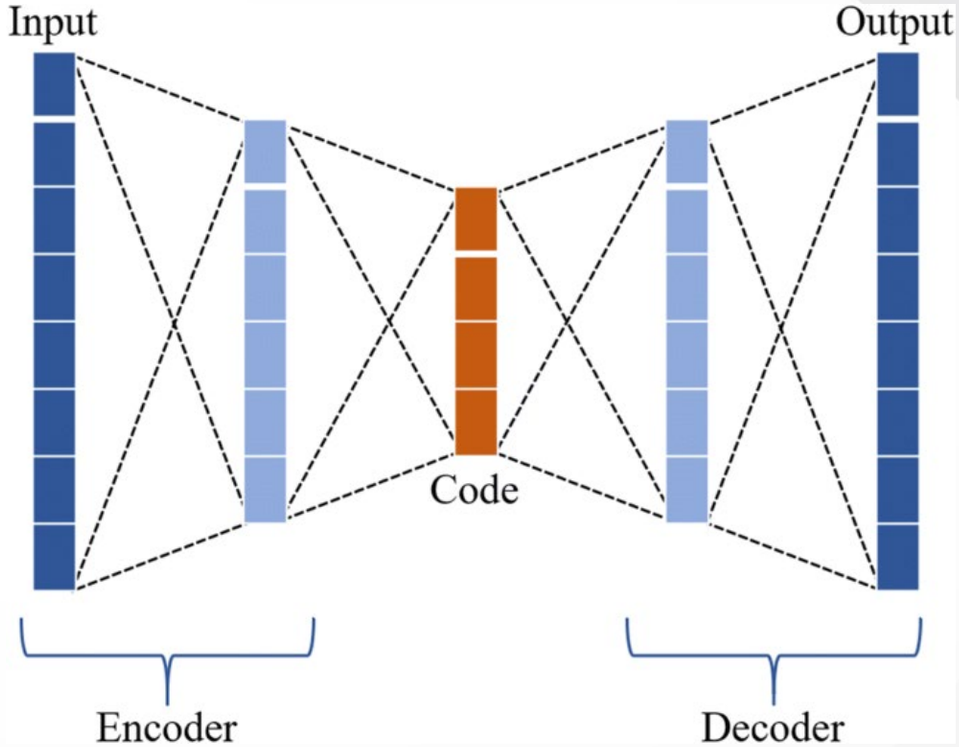
$$b_{i,k} = -\min W_{i,k} x_i$$

- Разобьём все сдвиги в архитектуре сети на типы I и II
- Фактически $\theta = W$

Санкт-Петербургский государственный университет

Метафора декодирования для выбора архитектуры сети





Санкт-Петербургский государственный университет

Оптимальный код. Бустинг

- Каждый слой сети – непрерывный канал передачи данных с шумом округления
- Скрытое представление – [M,N] код коррекции ошибок
  - Если шум не коррелирован с параметрами, то для достаточно большого N случайный код декодируется без ошибок с $P = 1 - \varepsilon$
- Каждый бит скрытого представления – слабый классификатор
  - Чем больше классов тем длиннее код можно построить в явном виде
- Унитарный (One-hot) код – бинарный классификатор один ко многим
  - Несбалансированный
  - Экспоненциальный рост с ростом числа классов
- Можно ли лучше?
  - Нейросети = деревья решений
  - Бустинг – построение сильного классификатора из слабых
  - Обучение нейросетевого полносвязного декодера = Нейробустинг

Выводы

- Основные практические достоинства нейросетей
  - Вычислительная эффективность вывода
  - Простота обучения
  - Возможность использования предобученных весов
- Основные практические недостатки нейросетей
  - Зависимость результата от обучающих данных
  - Необходимость сбора и разметки большого количества обучающих данных (в большинстве современных сценариев обучения)
  - Высокая вычислительная сложность обучения
  - Сложность проверки устойчивости и робастности решений

Санкт-Петербургский государственный университет

Дополнительная литература

[1] Aggarwal C. C. et al. Neural networks and deep learning //Springer. – 2023

[2] Moser S. M. Information theory, 6-th ed. //Lecture Notes. – 2018.

[3] David A. Patterson, John L. Hennessy. Computer Organization and Design RISC-V Edition The Hardware Software Interface, 2nd Edition // Elsevier Science. – 2020.