



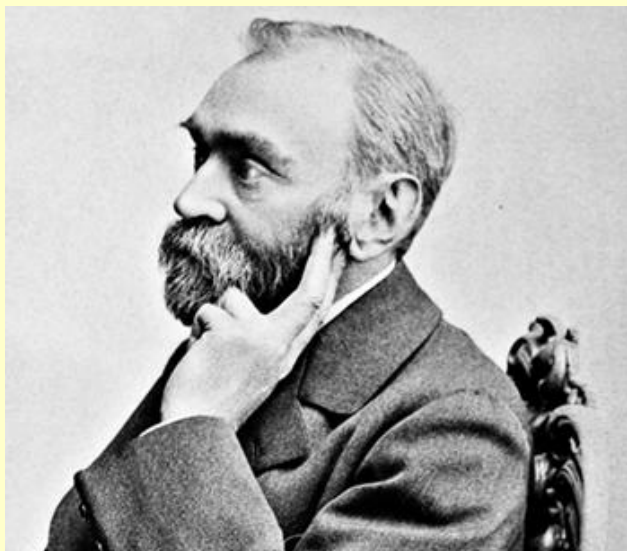
**Нобелевская премия по физике 2024:
почему её получили не физики,
или
Как физические принципы легли в основу
методов машинного обучения**

С.А. Доленко



Лаборатория адаптивных методов обработки данных НИИЯФ МГУ

АЛЬФРЕД НОБЕЛЬ И ПРЕМИЯ ЕГО ИМЕНИ



Альфред Бернхард Нобель

Родился: 21 октября 1833 года,
Стокгольм, Шведско-норвежская уния

Умер: 10 декабря 1896 года (63 года),
Сан-Ремо, Королевство Италия

Чем известен: 355 патентов
изобретатель динамита
учредитель Нобелевской премии

Нобелевская премия:

- Учреждена в 1895 году, вручается с 1901 года
- «Тем, кто в предыдущем году принесет наибольшую пользу человечеству» (сейчас это совсем не так)
- Сферы деятельности: физика, химия, медицина, литература, движение за мир. В 1968 году добавили экономику
- Вручается ежегодно в день смерти Нобеля (10 декабря)
- Не может присуждаться посмертно, может не присуждаться
- Сумма составляет 11 млн. шведских крон (~1 млн. долларов)

НОБЕЛЕВСКАЯ ПРЕМИЯ ПО ФИЗИКЕ 2024 ГОДА



NOBELPRISET I FYSIK 2024
THE NOBEL PRIZE IN PHYSICS 2024



KUNGL.
VETENSKAPS-
AKADEMIEN

THE ROYAL SWEDISH ACADEMY OF SCIENCES



Photo: Mary Whitham

John J. Hopfield

Princeton University, NJ, USA



Photo: University of Toronto

Geoffrey E. Hinton

University of Toronto, Canada

"för grundläggande upptäckter och uppfinningar som möjliggör maskininlärning med artificiella neuronnätverk"

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

#NobelPrize

THE
NOBEL
PRIZE

**«За основополагающие открытия и изобретения,
сделавшие возможным машинное обучение
с помощью искусственных нейронных сетей»**

ОНИ НЕ ФИЗИКИ ?

John Joseph Hopfield

- Родился: 15.07.1933, Chicago, США
- Образование: Cornell University
- PhD: 1958, физика, Cornell University
- Области деятельности: физика, биология, молекулярная биология, нейробиология, психология, информатика (машинное обучение)
- Место работы: Bell Labs; **Princeton University**; University of California, Berkeley; California Institute of Technology



J.J.Hopfield, G.E.Hinton

© The Royal Swedish Academy of Sciences

Geoffrey Everest Hinton

- Родился: 06.12.1947, Wimbledon, London, England
- Образование: King's College, Cambridge (1970)
- PhD: 1978, искусственный интеллект, University of Edinburgh
- Области деятельности: экспериментальная психология, кибернетика, информатика (машинное обучение), искусственный интеллект
- Место работы: **University of Toronto**; Google; Carnegie Mellon University; University College, London; University of California, San Diego

Премия Тьюринга
(2018)

МАШИННОЕ ОБУЧЕНИЕ

Машинное обучение (machine learning):

обучение машин, а не обучение при помощи машин!!!



Основное направление
машинного обучения:
построение
аппроксимационных
моделей на примерах

Изображение сгенерировано
нейронной сетью **mid journey**

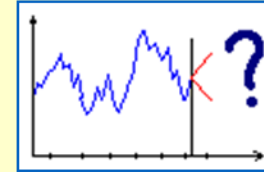
АППРОКСИМАЦИОННЫЕ МЕТОДЫ И ИХ МЕСТО

- Большинство рассматриваемых задач представляют собой задачи построения модели – **функционального отображения**
- Наилучший вариант – когда функцию можно задать **аналитически** на основе физической модели объекта
- Если аналитическое решение невозможно – пытаемся решить задачу **численно** на основе математической модели
- Если содержательных моделей нет – остается построение **аппроксимационной** модели на основе имеющихся примеров
- Аппроксимационная модель = **разложение** искомого отображения
- Таким образом, «**обучение на примерах**» = построение аппроксимационной модели, т.е. «**черного ящика**»
- Если есть возможность, ящик лучше **покрасить**, т.е. использовать имеющуюся априорную информацию

ТИПОЛОГИЯ ЗАДАЧ, РЕШАЕМЫХ МЕТОДАМИ МО

❑ Регрессия (оценка значения величины)

- Количественная оценка $Y = 378.14 \pm 0.21$
- Прогнозирование временных рядов
- Аппроксимация зависимостей

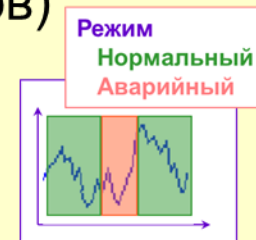


$$Y = F(X_1, X_2, \dots, X_n)$$

❑ Классификация (распознавание образов)

- Бинарная
- Многоклассовая
- Многометочная

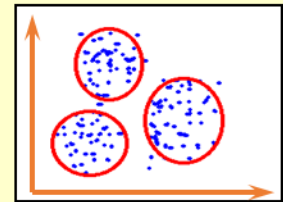
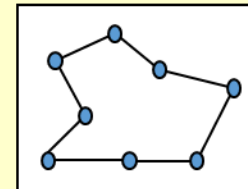
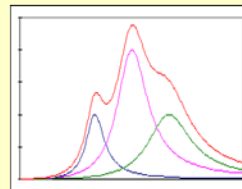
А Б В Г Д



1 2 3 4 5 6

❑ Оптимизация (поиск оптимальных значений / комбинаций)

- Поиск оптимальных значений
- Комбинаторная оптимизация



❑ Кластеризация (разбиение данных на группы)

Границы между типами задач могут быть размыты, одну и ту же задачу можно ставить как задачи разных типов

МАШИННОЕ ОБУЧЕНИЕ И ГЛУБОКОЕ ОБУЧЕНИЕ

- Как правило, решение задачи машинного обучения состоит из двух этапов:
 - Формирование пространства признаков, наиболее адекватно описывающих данные для решения рассматриваемой задачи
 - Собственно обучение алгоритмов с использованием сформированных признаков в качестве входных
- Глубокая нейронная сеть с каждым следующим слоем формирует всё более сложные композитные признаки, причём это происходит непосредственно в процессе обучения
- В последние годы глубоким обучением часто стали называть любые процессы обучения, в которых оптимальные признаки формируются в процессе обучения
- Это можно сделать не только с помощью персептронов

ТРИ ВОЛНЫ ИНТЕРЕСА К НЕЙРОННЫМ СЕТЯМ

Краткая историческая справка

- Формальный нейрон (McCulloch, Pitts) (1943)
- **Первая волна (1957-1969)**
 - Теорема Колмогорова о представлении функции n переменных (1957)
 - Персептрон (Rosenblatt) (1958)
 - **Minsky, Papert «Персептроны» (1969)**
- **Вторая волна (1982 – середина 1990-х)**
 - **Нейронная сеть Хопфилда (Hopfield) (1982)**
 - **Машина Больцмана (1985) (Hinton, Sejnowsky)**
 - **Алгоритм обратного распространения ошибки (1986) (Rumelhart, Hinton, Williams)**
 - Теорема об универсальной аппроксимации персептроном (1989) (Hornik, Stinchcombe, White; Cybenko; Funahashi)
 - Свёрточные нейронные сети (1989) (LeCun)
 - Обучение с подкреплением (Q-Learning)

ТРИ ВОЛНЫ ИНТЕРЕСА К НЕЙРОННЫМ СЕТЯМ

Краткая историческая справка

- **Третья волна (середина 2000-х – наши дни)**
 - Глубокие нейронные сети
 - Глубокие свёрточные нейронные сети
 - Рекуррентные нейронные сети с долгой краткосрочной памятью (LSTM)
 - Генеративные нейросетевые архитектуры
(генеративные состязательные сети, вариационные автоэнкодеры и др.)
 - Обучение с подкреплением (на глубоких сетях)
 - **Механизм внимания (трансформеры)**
 - **Большие языковые модели**

НЕКОТОРЫЕ УСПЕХИ ТРЕТЬЕЙ ВОЛНЫ

- ❑ Решена задача **дикторонезависимого** распознавания **речи**
- ❑ Решена задача **идентификации** (классификации) объектов на **изображениях** и в видеопотоке
- ❑ Решена задача **самообучения сложным играм** на основе знания лишь правил игры (шахматы, сёги, го)
- ❑ Достигнуты существенные успехи в машинном **анализе текстов**
- ❑ Достигнуты существенные успехи в машинном **перевode текстов**
- ❑ Достигнуты существенные успехи в **генерации новых данных**
- ❑ Достигнуты заметные успехи в общении **на естественном языке**

Полученные результаты основаны на доступности

- Существенно возросших **вычислительных мощностей** («закон Мура»)
- Практически неограниченного **объёма данных** некоторых типов
- Серьёзных **инвестиций**, связанных с уже достигнутыми успехами

Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2554–2558, April 1982
Biophysics

Neural networks and physical systems with emergent collective computational abilities

(associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)

J. J. HOPFIELD

Division of Chemistry and Biology, California Institute of Technology, Pasadena, California 91125; and Bell Laboratories, Murray Hill, New Jersey 07974

Contributed by John J. Hopfield, January 15, 1982

ABSTRACT Computational properties of use to biological organisms or to the construction of computers can emerge as collective properties of systems having a large number of simple equivalent components (or neurons). The physical meaning of content-addressable memory is described by an appropriate phase space flow of the state of a system. A model of such a system is given, based on aspects of neurobiology but readily adapted to integrated circuits. The collective properties of this model produce a content-addressable memory which correctly yields an entire memory from any subpart of sufficient size. The algorithm for the time evolution of the state of the system is based on asynchronous parallel processing. Additional emergent collective properties include some capacity for generalization, familiarity recognition, categorization, error correction, and time sequence retention. The collective properties are only weakly sensitive to details of the modeling or the failure of individual devices.

Given the dynamical electrochemical properties of neurons and their interconnections (synapses), we readily understand schemes that use a few neurons to obtain elementary useful biological behavior (1–3). Our understanding of such simple circuits in

calized content-addressable memory or categorizer using extensive asynchronous parallel processing.

The general content-addressable memory of a physical system

Suppose that an item stored in memory is “H. A. Kramers & G. H. Wannier *Phys. Rev.* **60**, 252 (1941).” A general content-addressable memory would be capable of retrieving this entire memory item on the basis of sufficient partial information. The input “& Wannier, (1941)” might suffice. An ideal memory could deal with errors and retrieve this reference even from the input “Vannier, (1941)”. In computers, only relatively simple forms of content-addressable memory have been made in hardware (10, 11). Sophisticated ideas like error correction in accessing information are usually introduced as software (10).

There are classes of physical systems whose spontaneous behavior can be used as a form of general (and error-correcting) content-addressable memory. Consider the time evolution of a physical system that can be described by a set of general coordinates. A point in state space then represents the instantaneous condition of the system. This state space may be either

СИСТЕМЫ КОЛЛЕКТИВНОГО ВЗАИМОДЕЙСТВИЯ. СИСТЕМА АТОМОВ

Энергия системы взаимодействующих магнитных моментов

$$E(s) = - \sum_i \sum_j M_{ij} s_i s_j$$

M_{ij} Коэффициент обменного взаимодействия

1. Ферромагнетики – ($M_{ij} > 0$) спины стремятся сориентироваться параллельно
2. Антиферромагнетики – ($M_{ij} < 0$) спины соседних атомов направлены в противоположные стороны
3. Спиновые Стекла – связи между атомами носят случайный характер

НЕЙРОННАЯ СЕТЬ ХОПФИЛДА

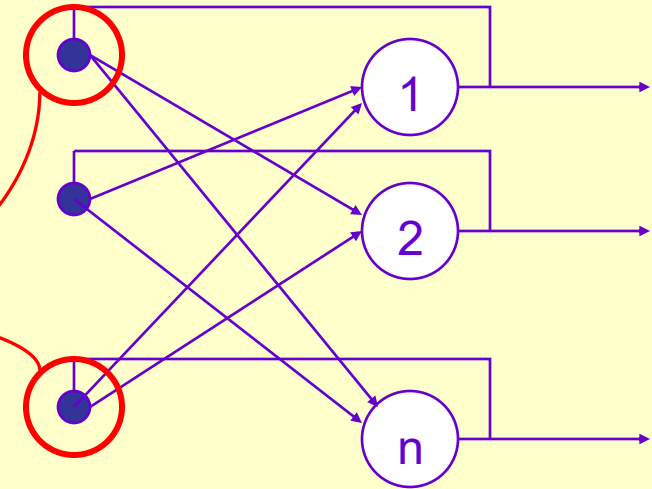
Является **рекуррентной**,
т.е. содержит обратные связи и может
эволюционировать с течением времени.

Каждый нейрон осуществляет
нелинейное преобразование

$$\xi_i(t+1) = \Phi\left(\sum_{\substack{j=1 \\ j \neq i}}^n w_{ij} \xi_j(t) - T_i\right)$$

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{pmatrix}$$

Матрица связей



	Система атомов	Нейронная сеть
Структура	Система связанных диполей	Система связанных нейронов
Энергия	$E = -\sum_i \sum_j M_{ij} S_i S_j$	$E = -\sum_i \sum_j w_{ij} \xi_i \xi_j$

ХАРАКТЕРИСТИКИ СЕТЕЙ ХОПФИЛДА

В любой сети Хопфилда матрица связей симметрична и имеет нулевую главную диагональ

$$W = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{pmatrix}$$

Нейроны: - бинарные

- непрерывные

Динамика: - асинхронная

- последовательная

- параллельная

1. $\forall (i, j) \quad w_{ij} = w_{ji}$

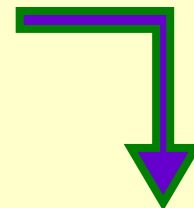
2. $\forall i \quad w_{ii} = 0$

3. $\forall i \quad \xi_i \in [\xi_{MIN}, \xi_{MAX}]$

Эти свойства обеспечивают сходимость эволюции

Такие сети, подобно спиновым стеклам, имеют множество аттракторов, которым соответствуют минимумы энергии

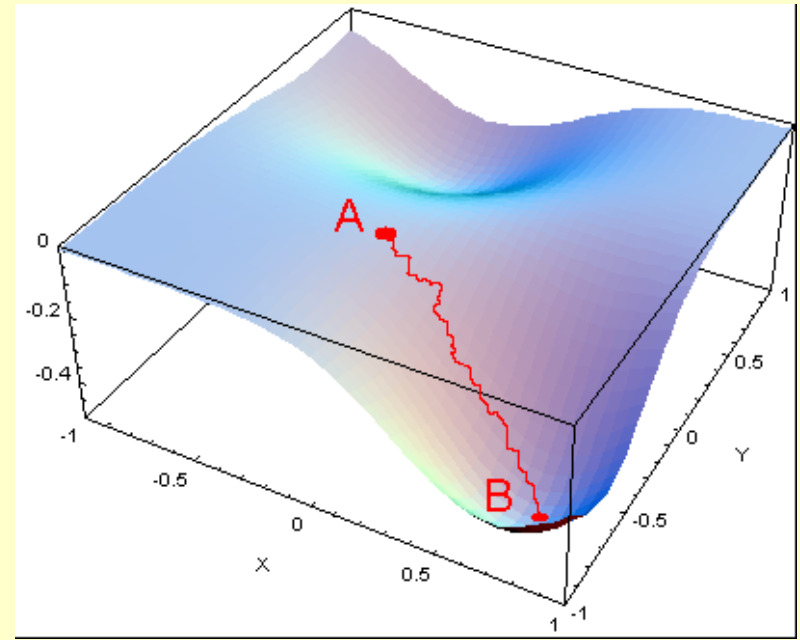
Сети Хопфилда могут использоваться для реализации ассоциативной памяти



ЭНЕРГЕТИЧЕСКИЙ ЛАНДШАФТ

Сети Хопфилда,
будучи «отпущенными»
из произвольного
начального состояния,
эволюционируют к состоянию
с минимальной энергией

Энергетический ландшафт
формируется только
матрицей связей между нейронами



Задача

Уметь задавать форму энергетического ландшафта
путем настройки матрицы связей.

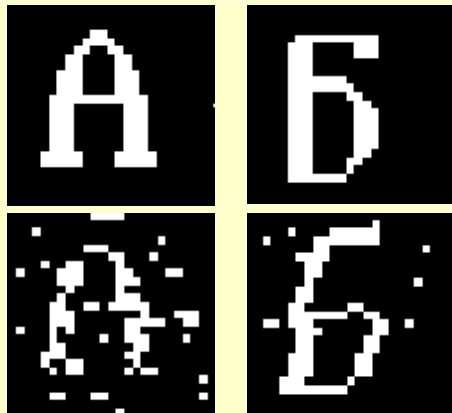
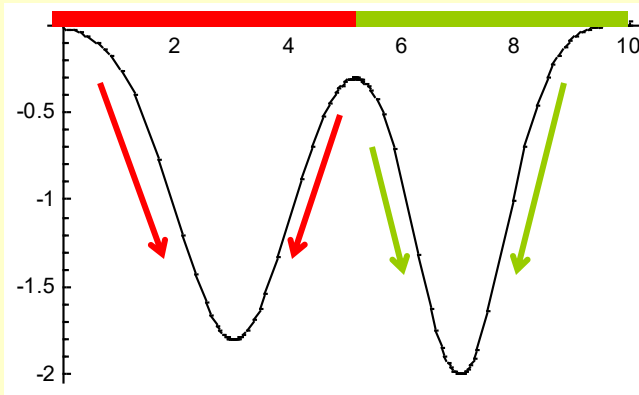
Желательные состояния должны обладать минимальной энергией₁₆

ПРАВИЛО ХЕББА

$$\mathbf{Y} = \begin{bmatrix} \xi^1 \\ \xi^2 \\ \dots \\ \xi^L \end{bmatrix} = \begin{bmatrix} \xi_1^1 & \xi_2^1 & \dots & \xi_n^1 \\ \xi_1^2 & \xi_2^2 & \dots & \xi_n^2 \\ \dots & \dots & \dots & \dots \\ \xi_1^L & \xi_2^L & \dots & \xi_n^L \end{bmatrix}$$

$$\mathbf{W} = \mathbf{Y}^T \mathbf{Y} - \mathbf{I}$$

В матрице $\mathbf{Y}$: строки – примеры,
столбцы – признаки



После обучения нейронной сети
входные образы $\xi^l, l \in \{1, \dots, L\}$
образуют впадины, соответствующие
набору хранимой информации

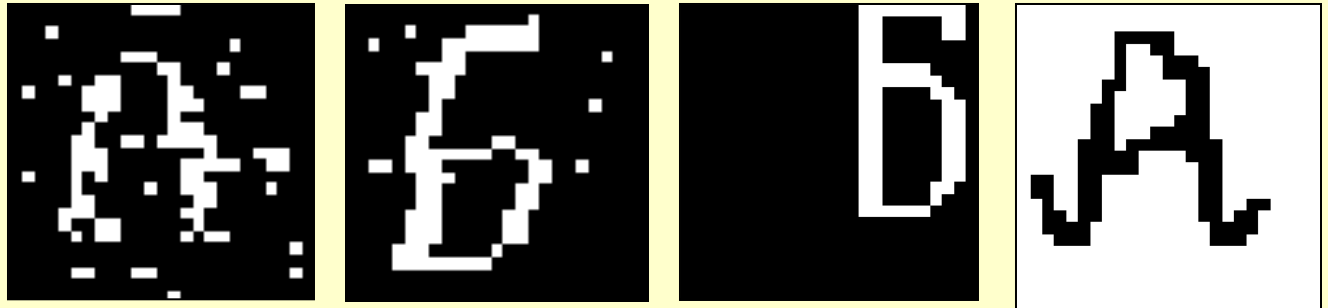
- Образам, запомненным сетью, соответствуют локальные минимумы
- Зашумленным образам соответствуют точки в бассейнах притяжений хранимых в памяти образов

“РАСПОЗНАВАНИЕ” ОБРАЗОВ

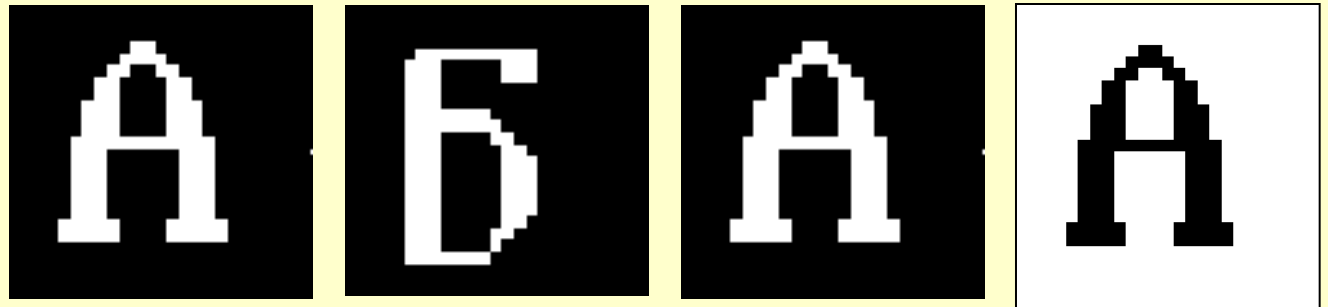
Образы в
памяти сети



Образы,
подаваемые
на вход



Образы,
восстанов-
ленные сетью



ЕМКОСТЬ ПАМЯТИ



Такие комбинации из
эталонных образов
называются *химерами*



Объем памяти сети для устойчивой работы M_0
зависит от количества нейронов n

$$M_0 = \frac{n}{4 \ln(n)}$$

$$M_{\max} \approx 0.138 \cdot n$$

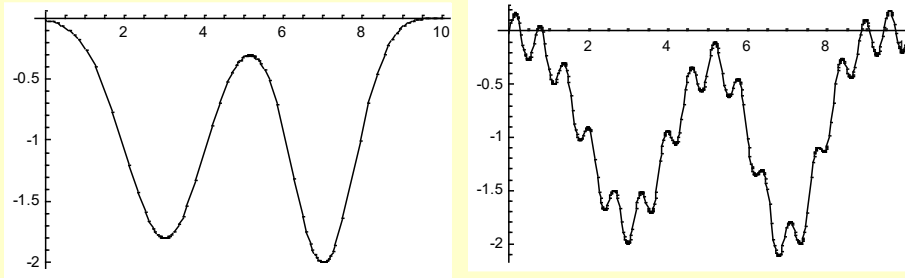
При $M > M_0$ и Хеббовской матрице связей происходит
“катастрофа памяти” – забывание всех образов

При использовании квази-Хеббовской матрицы
со своим весом для каждого примера
забываются примеры с весом ниже порога
(Я.М.Карандашев, Б.В.Крыжановский, Л.Б.Литинский, 2010)

Лучше всего запоминаются линейно независимые
и ортогональные векторы

МАШИНА БОЛЬЦМАНА

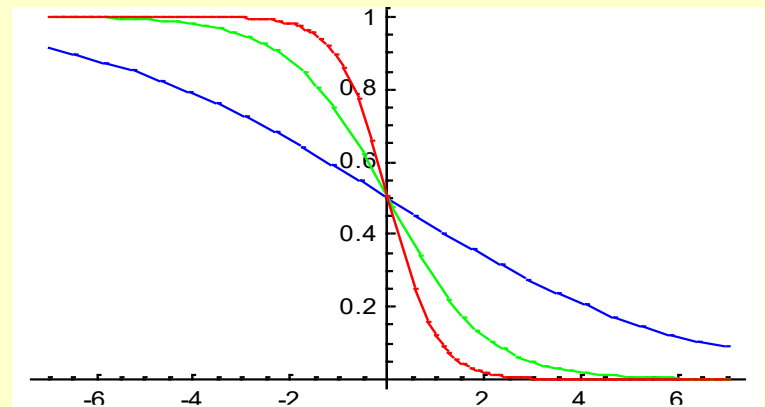
(1985) Hinton, Sejnowsky



Классическая сеть Хопфилда эволюционирует в ближайший энергетический минимум, *всегда уменьшая* свою энергию

В Машине Больцмана вероятность перехода в новое состояние

$$P_i(\xi_i) = \frac{1}{1 + e^{\left(\frac{\Delta E_i}{T}\right)}}$$

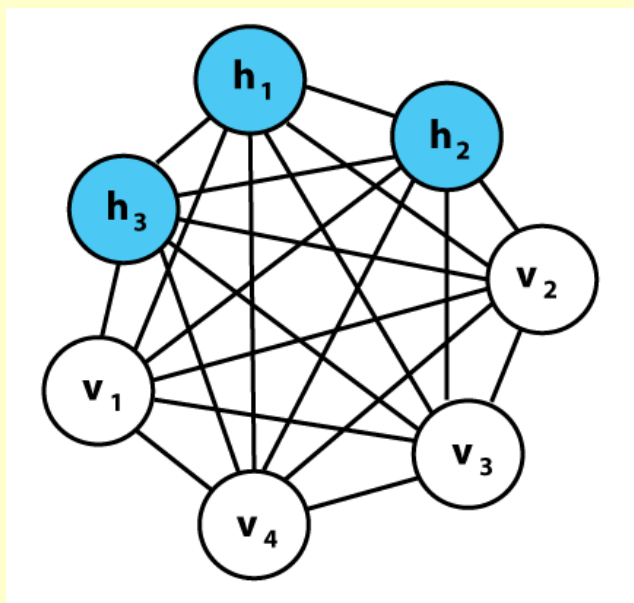


1. Случайно выбирается нейрон, для него вычисляем новое значение $\xi_i = random(0, 1)$
2. Если $\Delta E_i \leq 0$, фиксируем новое значение; если $\Delta E_i > 0$, то новое значение принимается с вероятностью $P_i(\xi_i)$
3. $T(t) = \frac{T_0}{1 + \log(t)}$

ОГРАНИЧЕННАЯ МАШИНА БОЛЬЦМАНА

Машина Больцмана:

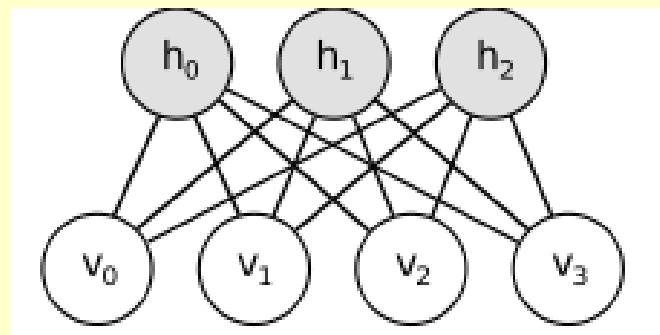
Скрытый
слой



Видимый
слой

Ограниченная
машина Больцмана:

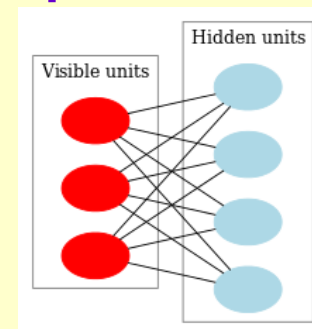
Скрытый
слой



Видимый
слой

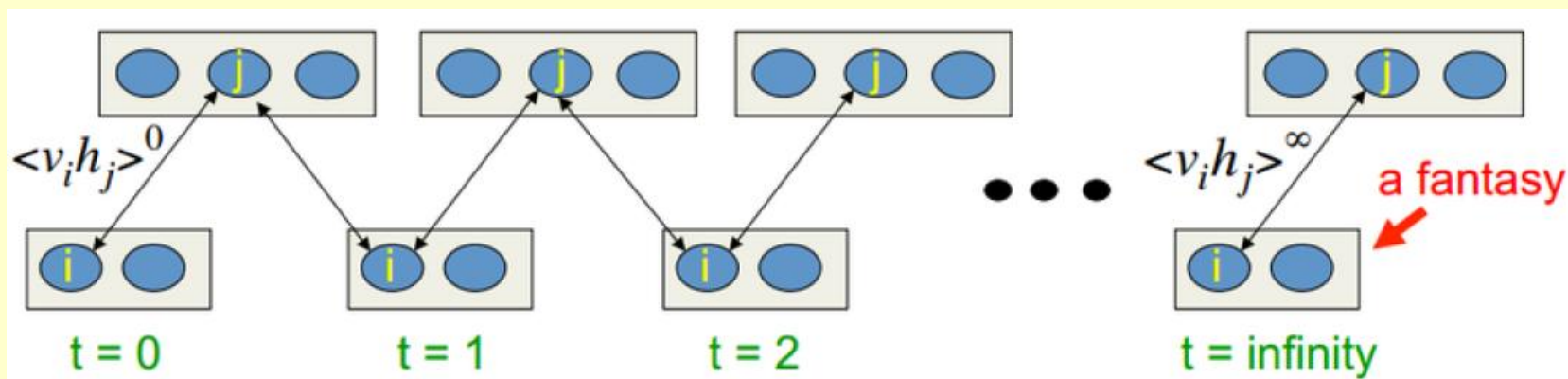
ОГРАНИЧЕННАЯ МАШИНА БОЛЬЦМАНА

- Машина Больцмана = сеть Хопфилда с ненулевой вероятностью перехода в состояние с большей энергией (распределение по Гиббсу)
- Ограниченная машина Больцмана – нейроны делятся на две группы, «видимые» и «скрытые»
- На основании состояния видимых нейронов оценивается распределение вероятностей состояний скрытых нейронов, и наоборот (Р. Хинтон); процедура быстро сходится



h

v

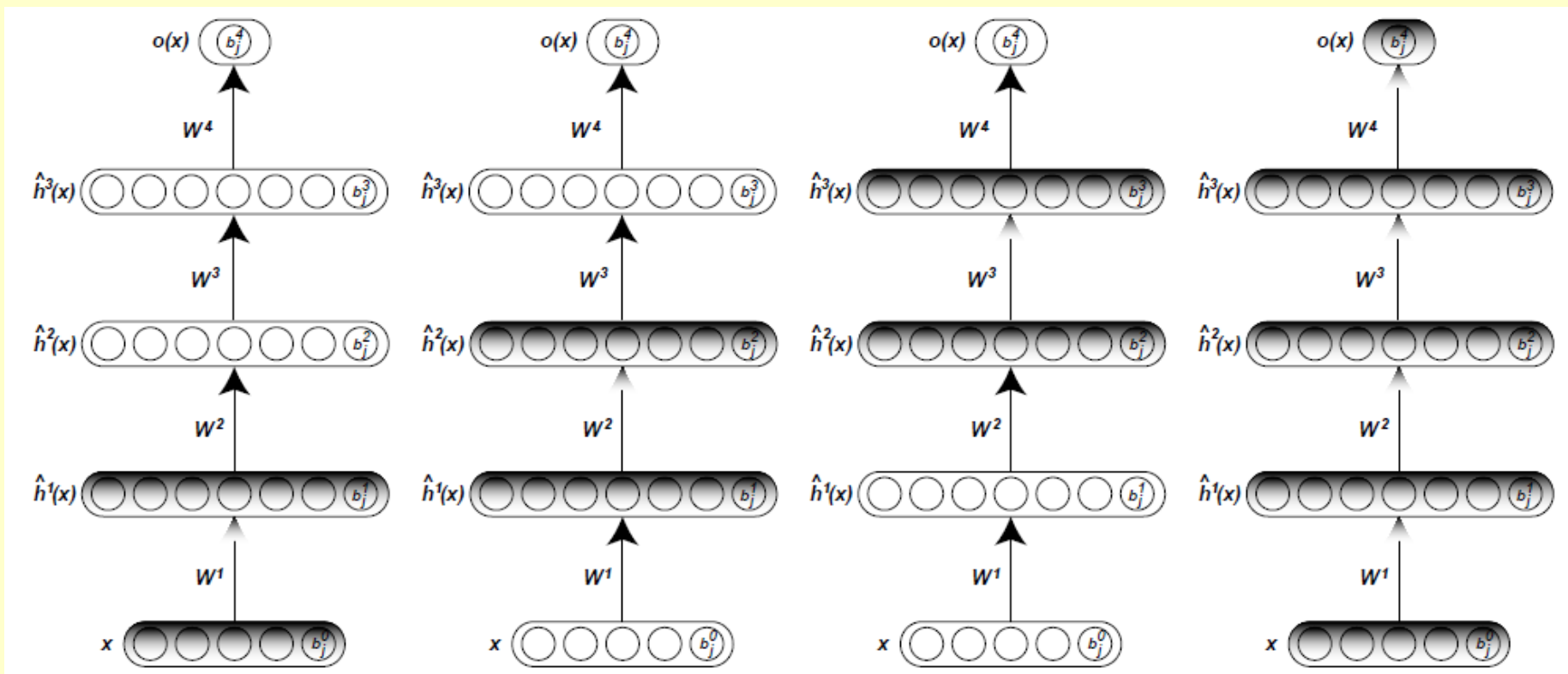


Алгоритм контрастного расхождения (contrastive divergence, CD)

ГЛУБОКИЕ НЕЙРОННЫЕ СЕТИ: ПРОБЛЕМЫ ОБУЧЕНИЯ МНОГОСЛОЙНЫХ СЕТЕЙ

- Рассмотрим **многослойный персептрон**, обучаемый по методу обратного распространения ошибки.
- Для скрытого слоя формула для подстройки весов **линейна относительно градиента**.
- При большом количестве слоёв (иногда их бывает до 30) возможно **затухание градиента**, если значение весов или производной в точке мало.
- Один из способов борьбы – увеличение **скорости обучения** по мере продвижения от выходного слоя к входному
- Однако чаще используется **предобучение** многослойных сетей
 - Автоэнкодеры (autoencoders)
 - **Ограниченная машина Больцмана**

ПРЕДОБУЧЕНИЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ: ОГРАНИЧЕННАЯ МАШИНА БОЛЬЦМАНА

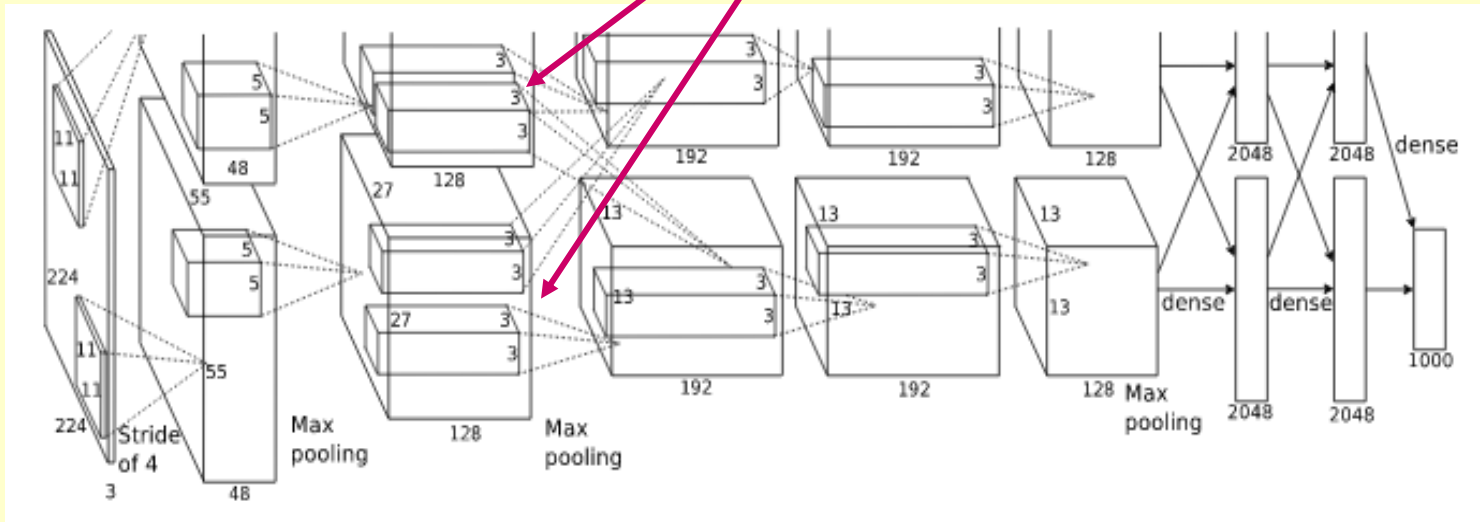


Послойное предобучение – подаем состояние предыдущего слоя на **видимые** нейроны ограниченной машины Больцмана, снимаем распределение **со скрытых** нейронов после применения алгоритма контрастного расхождения

ПРИМЕР УСПЕШНОЙ АРХИТЕКТУРЫ ГЛУБОКИХ СВЁРТОЧНЫХ НЕЙРОСЕТЕЙ: ALEXNET

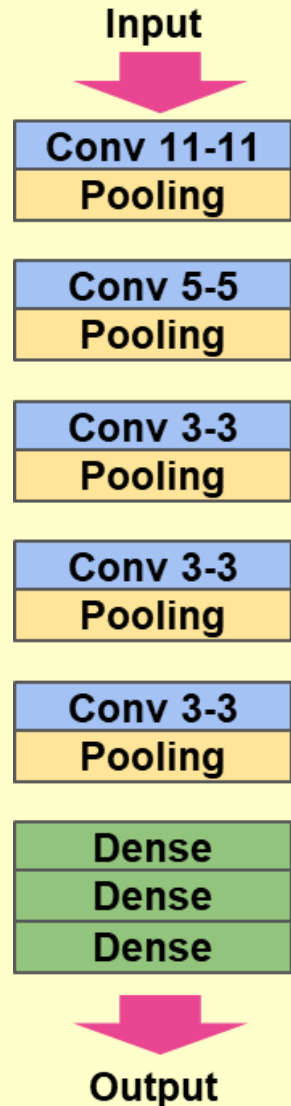
Krizhevsky, Sutskever, Hinton (2012)

Разделенная структура:
обучалась на 2х видеокартах
gtx 580 3gb 6 дней (90 эпох)



- Max pooling
- Relu, Dropout
- 60 000 000 параметров
- 12 000 000 примеров
- Image augmentation

Результат на ImageNet:
Ошибка top-5: 15.2%
Результат 2го места
того же года: 26.2%



ВЫБОРОЧНОЕ ПРОРЕЖИВАНИЕ (DROPOUT)

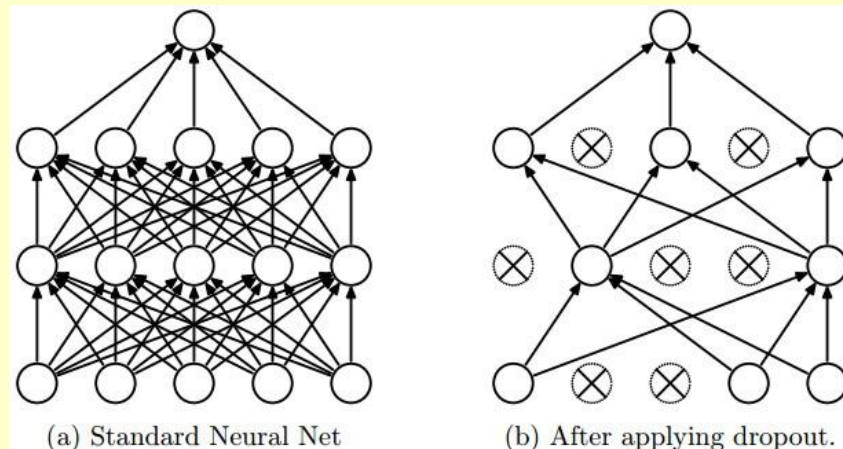
- Обучение сети производится градиентным спуском со **случайным выбором** примера (2014)
- Для каждого примера **временно отключаем** каждый нейрон сети и все его связи с вероятностью p
- Обратное распространение ошибки и градиентный спуск прodelываются **по прореженной** сети
- Затем все отключённые нейроны и связи **включаются обратно**
- При применении выход каждого нейрона домножается на $(1 - p)$

Таким образом, одновременно обучаются 2^N возможных архитектур сети (N – количество нейронов)

Получаемый ответ – результат фактического усреднения ответов 2^N прореженных сетей

Одновременное использование регуляризации и комитета (ансамбля)

N.Srivastava, **G.Hinton**, A.Krizhevsky,
I.Sutskever, R.Salakhutdinov.
J. of Machine Learning Research,
2014, V.15, pp.1929-1958.



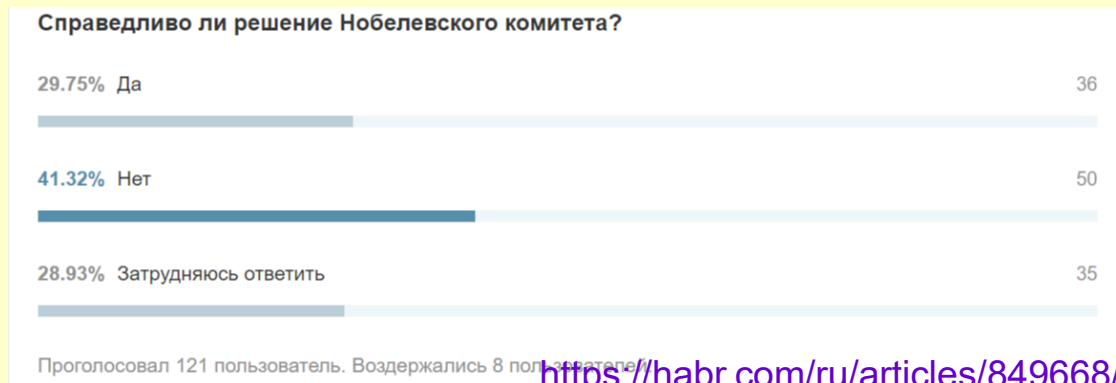
- **Капсульные нейронные сети (2017)**
 - Вместо отдельных нейронов используются капсулы – группы нейронов, представляющие свойства объекта (положение, ориентацию, размер, скорость и др.)
 - Нейрон на выходе выдает скаляр, капсула – вектор
 - Динамическая маршрутизация – выборочное распространение сигнала между капсулами разных уровней
- **Модель Transformer Foundation (2023)** (обработка естественного языка)
- **Опасности и риски, связанные с развитием ИИ**
 - Ушел из Google (2023), чтобы иметь возможность свободно высказываться на эту тему
 - Концепция «Материнского инстинкта» (2025): закладывать в архитектуру ИИ фундаментальное стремление заботиться о людях, аналогичное биологической привязанности матери к ребёнку.

И ВСЁ-ТАКИ: ЗА ЧТО, И ПРИ ЧЁМ ЗДЕСЬ ФИЗИКА?

- Сеть Хопфилда – использовала в качестве прототипа принципы и идеи, заимствованные **из физики**
- Машина Больцмана – дальнейшее проникновение и использование **физических** принципов в построении алгоритма обработки информации
- Широкое применение методов машинного обучения для обработки данных **в физике**
- Результаты работ, открытий и изобретений номинантов приносят и уже принесли «пользу человечеству»

Справедливо ли это?

Решайте сами...



СПИСОК ЛИТЕРАТУРЫ

1. Альфред Нобель: биография пацифиста и изобретателя динамита.
<https://trends.rbc.ru/trends/social/651ea4cb9a794749c327f621>
2. The Royal Swedish Academy of Sciences. The Nobel Prize in Physics 2024. Press Release, 8 October 2024.
3. Nobel Prize in physics awarded to 2 scientists for discoveries that enable machine learning.
<https://phys.org/news/2024-10-nobel-prize-physics-awarded-discoveries.html>
4. J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA, V.79, pp.2554-2558 (1982).
5. D.E. Rumelhart, G.E. Hinton, R.J. Williams. Learning representations by back-propagating errors. Nature, V.323, pp.533-536 (1986).
6. N.Srivastava, G.Hinton, A.Krizhevsky, I.Sutskever, R.Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. J. of Machine Learning Research, V.15, pp.1929-1958 (2014).



NOBELPRISET I FYSIK 2024
THE NOBEL PRIZE IN PHYSICS 2024



KUNGL.
VETENSKAPS-
AKADEMIEN

THE ROYAL SWEDISH ACADEMY OF SCIENCES

Спасибо за внимание !

