

Od neuronu do biliona dolarów

Opowieść w dwunastu rozdziałach o tym, jak sztuczna inteligencja doszła tam, gdzie jest, i kto na tym zarabia

BAZYLI ZYGAN

Od neuronu do biliona dolarów

Opowieść w dwunastu rozdziałach o tym, jak sztuczna inteligencja doszła tam, gdzie jest, i kto na tym zarabia

Copyright © 2026 Bazyli Zygan

Tekst i rysunki © Bazyli Zygan. Wszelkie prawa zastrzeżone. Żadna część tekstu ani rysunków nie może być powielana ani rozpowszechniana bez pisemnej zgody autora, z wyjątkiem dozwolonego użytku przewidzianego prawem.

Fotografie i reprodukcje są dziełami innych autorów i pozostają na swoich licencjach (Creative Commons, domena publiczna i inne). Autora, źródło, licencję i wprowadzone zmiany każdej z nich podaje spis „Ilustracje” (s. 327), a zasady korzystania z nich – „Licencje i informacje prawne” (s. 339). Źródła, na których opiera się tekst, zebrano w bibliografii (s. 315).

Liczby, ceny, wyceny i inne zmienne dane podajemy według stanu na 10 października 2026 roku, chyba że tekst mówi inaczej.

Książka ma charakter publicystyczny i edukacyjny. Nie jest poradą inwestycyjną ani rekomendacją kupna lub sprzedaży jakichkolwiek instrumentów finansowych.

Na stronie tytułowej: Krzem karty GeForce GTX 580 · Fritzchens Fritz · CC0.

Skład: Manrope i JetBrains Mono (SIL Open Font License 1.1).

Spis treści

Wprowadzenie

6

Wstęp do opowieści w dwunastu rozdziałach o tym, jak sztuczna inteligencja zaszła tak daleko – opowieści, w której ktoś kupuje dwie karty graficzne z zupełnie niewłaściwych powodów, a czternaście lat później sieć energetyczna zaczyna nerwowo mrugać.

CZĘŚĆ I: INTELIGENCJA Jak działa sztuczna inteligencja

01 Jak uczą się maszyny

11

W którym maszynie wielkości pokoju obiecuje się złote góry, jedna książka psuje zabawę całemu pokoleniu, a uczenie się okazuje się schodzeniem z góry we mgle.

02 Maszyny, które widzą

27

W którym kot ignoruje pokaz slajdów, pani profesor liczy zdjęcia, a dwie karty graficzne wreszcie znajdują porządne zajęcie.

03 Attention Is All You Need

43

W którym zdanie wędruje od okienka do okienka, aż zapomina, w jakiej sprawie przyszło, ośmioro ludzi z jednego korytarza postanawia, że koniec ze stanem w kolejce, a artykuł z tytułem ściągniętym z popowej piosenki przemeblowuje całą branżę.

04 Jak zbudować model językowy

64

W którym maszyna przystępuje do najdłuższego egzaminu w dziejach, zmyśla sześć wyroków, pobiera lekcje dobrych manier u czterdziestu zleceniobiorców i wystawia dwa rachunki.

CZĘŚĆ II: MOC OBLICZENIOWA I ENERGIA Warstwa fizyczna

05 Ekonomia krzemu

88

W którym firmę obmyśla się w przydrożnym barze przy kawie dolewanej bez końca, pięćdziesięciopięciolatek otwiera fabrykę bez własnych wyrobów, pewna maszyna przylatuje trzema samolotami, a cała kolejka płaci, jak się okazuje, w tym samym sklepie.

06 Prąd i obliczenia alternatywne

112

W którym elektrownię jądrową wskrzesza się po to, żeby odpowiadała na pytania, największa sieć energetyczna Ameryki pyta najmądrzejsze maszyny świata, czy nie zechciałyby przejść na diesla, a kilka osób proponuje, żeby rachować wodą, światłem albo czymś, co ma mózg sowy.

07 **Obrazy w ruchu** 139

W którym magnat kolejowy płaci za to, żeby się dowiedzieć, co koń robi z nogami, ktoś wpada w pubie na pewien pomysł, nauczyciel spod Hamburga spisuje pięć miliardów obrazków, papież sprawia sobie kurtkę, a najsłynniejszą na świecie maszynę do filmów wyłącza się, bo wychodzi za drogo.

08 **Poza Transformerem** 165

W którym berliński koń rachuje przed trzynastoma ekspertami, twierdzenie, które pomógł udowodnić dziewiętnastolatek, czeka na posadę sześćdziesiąt siedem lat, dwóch uczonych zakłada się o to, co będzie pierwszego stycznia 2027 roku, maszyny uczą się myśleć, zanim się odezwą, i przysyłają za to rachunek, a jedna z nich dostaje pod opiekę sklepik i sprzedaje wolfram ze stratą.

09 **Szok DeepSeeka** 195

W którym francuski ksiądz w chińskim miasteczku pisze do domu list o talerzach, fundusz hedgingowy z Hangzhou znajduje sobie hobby i zaczyna notować skrótami, zastrzeżenie, którego nikt nie czyta, w jeden dzień kosztuje producenta chipów 589 miliardów dolarów, profesor tłumaczy, co BMW ma wspólnego ze śmieciarką, a najcenniejszej tajemnicy epoki strzeże, jak się okazuje, regulamin.

10 **Chiny dziś** 228

W którym pewien człowiek – nic więcej o nim nie wiadomo – rozdaje najstarszą datowaną drukowaną książkę świata, słowo, którym opisywano jawajskie pola ryżowe, dojeżdża na rowerze do Biura Politycznego, szef wyszukiwarki nazywa darmowe oprogramowanie podatkiem od głupoty, po czym sam wypuszcza dziesięć darmowych modeli, reguła wymyślona na Stanfordzie w 1959 roku uczy chińskie maszyny skreślać, wojna cenowa ustaje, bo w sklepie zabrakło towaru, a giełda przez chwilę wycenia pewne laboratorium na tysiąc sześćset razy tyle, ile wynosi jego sprzedaż.

11 **Europa: na boisku czy na trybunach?** 262

W którym pewien optyk nie dostaje patentu na lunetę, bo nie da się jej utrzymać w tajemnicy, firma z przeciekającej szopy zostaje jedynym producentem maszyny, która w większości już istniała, trzej Francuzi dostają sto pięć milionów euro za siedem stron, orzeł, który ściśle biorąc orłem nie jest, wykluwa się ochotnikom na superkomputerze w Krakowie, a pewien kontynent odkłada przepisy, choć mówił, że ich nie odłoży.

12	Mapa łańcucha wartości	288
	W którym młody bostończyk sprzedaje na Antylach zamrożoną wodę ze stawu i trafia przez to do więzienia, filozof patrzy, jak jego staw odpływa do Kalkuty, przedsiębiorca z Tajpej rysuje uśmiech, kość pamięci niespodziewanie zarabia na każdym dolarze więcej niż najcenniejsza firma świata, wyszukiwarka pożyczka pieniądze na sto lat, browar otwiera bar tuż obok swoich najlepszych klientów, a książka kończy się tak, jak kończy się każda hossa: ktoś zaczyna liczyć.	
	Słowniczek	313
	Bibliografia	315
	Ilustracje	327
	Licencje i informacje prawne	339

Wprowadzenie

Wstęp do opowieści w dwunastu rozdziałach o tym, jak sztuczna inteligencja zaszła tak daleko – opowieści, w której ktoś kupuje dwie karty graficzne z zupełnie niewłaściwych powodów, a czternaście lat później sieć energetyczna zaczyna nerwowo mrugać.

OKŁADKA: CENTRA DANYCH POD ASHBURN W WIRGINII • THEODORE CHRISTOPHER • CCO

01

Dwa obrazy

Opowieści o sztucznej inteligencji powinny zaczynać się od robota, najlepiej z czerwonymi oczami i pretensjami do świata. Ta zaczyna się od zakupów.

W 2012 roku pewien doktorant w Toronto wytrenował sieć neuronową na dwóch kartach graficznych. Były to dwa GeForce'y GTX 580 – sprzęt zaprojektowany z wielką starannością i za niemałe pieniądze po to, żeby nastolatki mogły do siebie strzelać w wyższej rozdzielczości. W fabryce nikt nie myślał przy tym o sztucznej inteligencji. Karty nie wiedziały, że za chwilę zmienią świat. Karty graficzne rzadko o tym wiedzą.*



Obrazek pierwszy. GeForce GTX 580 trafił do sklepów pod koniec 2010 roku, kosztował mniej więcej pięćset dolarów i był przeznaczony dla ludzi, którzy traktowali gry wideo śmiertelnie poważnie. Całą robotę wykonały dwie takie karty.

FOT.: THESTRIKER • CC BY-SA 4.0

* Są to z natury przedmioty mało ciekawe świata. Pod tym i kilkoma innymi względami różnią się od ludzi, którzy je kupują.

Teraz obrazek drugi. Dziś firmy budują dla sztucznej inteligencji centra danych, które pożerają tyle prądu co małe miasto. Mają własne stacje transformatorowe. Mają nawet coś w rodzaju własnej pogody. W urzędach od planowania przestrzennego mówi się o nich nabożnym szeptem, jak kiedyś o katedrach i autostradach.



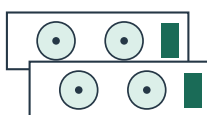
Obrazek drugi. Centrum danych Google'a w Council Bluffs w stanie Iowa na zdjęciu z 2017 roku – czyli, jak na tę branżę, z epoki kamienia łupanego. Te, które powstają teraz, są większe.

FOT.: CHAD DAVIS • CC BY 2.0

Od dwóch kart graficznych do apetytu na całą elektrownię – w jakieś czternaście lat. To nie trend. To lot z urwiska, tyle że w górę.

Czternaście lat apetytu

2012: SYPIALNIA W TORONTO



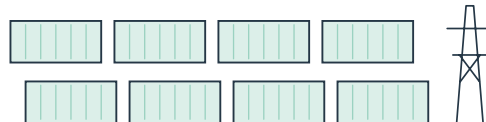
Dwie karty graficzne

Razem poniżej 500 watów.
Toster bierze więcej.

× 2,5 miliona

CZTERNAŚCIE LAT

2026: KAMPUS W TEKSASIE



Osiem budynków, jeden rachunek za prąd

W planach 1,2 gigawata. Według Oracle'a starczyłoby na milion domów.

Te same dwa obrazki, tym razem z licznikiem prądu. Na pełnych obrotach obie karty ciągnęły razem 488 watów. Jeden tylko kampus, który właśnie rośnie w Abilene w Teksasie, zaplanowano na 1,2 gigawata. Mało które hobby aż tak wymknęło się spod kontroli.

Cała ta opowieść krąży wokół jednego pytania o przepaść między tymi dwoma obrazkami: **co się wydarzyło po drodze i kto spija z tego śmietankę?** Pierwsza połowa to historia pomysłów, upartych ludzi i mnożenia na skalę przemysłową. Druga jest o pieniądzach, bo to w nich toczy się dziś ta historia, a płacimy za nią wszyscy.

02

Po co wiedzieć, na czym polega sztuczka

Są takie rozmowy – na zebraniu, przy kolacji, w wieczornych wiadomościach – w których ktoś mówi „model bazowy” i wszyscy kiwają głowami. Potem ktoś mówi „koszty inferencji” i wszyscy kiwają znowu, odrobinę szybciej. Kiedy pada słowo „parametry”, towarzystwo kiwa już równo jak pieski na tylnej półce samochodu i nikt nie śmie przestać, bo jeszcze ktoś go o coś zapyta.*

Chcielibyśmy być w takiej rozmowie tymi, którzy naprawdę wiedzą, o co chodzi. Nie po to, żeby te maszyny budować – to zajęcie dla miłośników algebry liniowej i najlepiej zostawić ich z nią sam na sam – tylko po to, żeby umieć je ocenić: czy są warte tego, co się na nie wydaje, kto za nie płaci i kto na nich zarabia. Trudno ocenić coś, czego nie umie się sobie wyobrazić. Jeszcze trudniej zauważyć, że jego rachunek ekonomiczny zaraz się zmienia – a w tej branży zmienia się mniej więcej co osiemnaście miesięcy, zwykle we wtorek.

Stąd dwanaście rozdziałów. Opowieść zaczyna się od pojedynczego sztucznego neuronu, który okazuje się mniej więcej tak skomplikowany jak członek komitetu inwestycyjnego.† Kończy się na chipach, elektrowniach i geopolityce, czyli na sporze o to, komu wolno co kupić.

Narrator tych rozdziałów zawodowo zajmuje się pieniędzmi, dlatego co chwila zaglądają do nich komitety inwestycyjne, transakcje i wnioski kredytowe. Czytelnik wcale nie musi. Pieniądze to po prostu najszybszy sposób, żeby zobaczyć, kto w tej grze wygrywa, a narrator obiecuje za każdym razem pokazać, jak liczył.

Matematyka nie będzie potrzebna. Potrzebna będzie ciekawość i odwaga, żeby zadać pytanie, które wszystkim na sali chodzi po głowie, choć każdy po cichu liczy, że zada je ktoś inny.

03

Trzy wątki

Dwanaście rozdziałów to spory kłębek. Przewijają się przez niego trzy wątki, a kto się ich trzyma, ten nie zabłądzi.

Architektura. Każdy przełom znosi ograniczenia poprzedniego. Historia sztucznej inteligencji to w pierwszym przybliżeniu długi korowód mądrych ludzi, których zirykował poprzedni mądry pomysł.

* Nazywa się to konsensusem.

† Bez urazy dla żadnej ze stron.

Dane, moc obliczeniowa i energia. Skala to paliwo. Chipy i prąd wyznaczają granice. Prędszej czy później każdy elegancki pomysł w tej dziedzinie musi stanąć w kolejce po sprzęt – a przed nim zawsze stoi jakaś większa firma.

Kto spija śmietankę? Ten wątek obchodzi każdego, kto ma w tym interesie własne pieniądze – a przez fundusze emerytalne, rachunki za prąd i podatki ma je w nim większość z nas. Technologia może zmienić świat, a wynalazcom przynieść grosze – o czym mogliby opowiedzieć wynalazcy niemal wszystkiego. Dlatego każdy rozdział kończy się rachunkiem: gdzie jest marża, kogo chroni solidny mur i czyj mur przy bliższych oględzinach okazuje się płotkiem.

04

Mapa

Dwanaście rozdziałów dzieli się na cztery części. Pierwsza wyjaśnia, jak to coś myśli, druga – czym się żywi, trzecia – dokąd zmierza, a czwarta – kto się o nie bije.

01 Jak uczą się maszyny

Neurony, warstwy, spadek gradientu i to, dlaczego sztuczna inteligencja czekała na swoją kolej trzydzieści lat.

02 Maszyny, które widzą

Sieci konwolucyjne i przełomowy rok 2012 w konkursie ImageNet, kiedy karty do gier weszły do gry o sztuczną inteligencję.

03 Attention Is All You Need

Od sieci rekurencyjnych do Transformera, architektury skrojonej na miarę sprzętu, który liczy równolegle.

04 Jak zbudować model językowy

Pretrening, prawa skalowania i RLHF. Czego naprawdę trzeba, żeby model wytrenować, a czego, żeby potem utrzymać go w ruchu.

05 Ekonomia krzemu

Dlaczego GPU rządzą, dlaczego tyle kosztują i kto trzyma w rękę łańcuch dostaw.

06 Prąd i obliczenia alternatywne

Ile prądu zjada sztuczna inteligencja i kto chce to zmienić: układy analogowe, fotoniczne i neuromorficzne.

07 Obrazy w ruchu

GAN-y, dyfuzja i Stable Diffusion, a także to, dlaczego wideo to jak dotąd najtwardszy orzech do zgryzienia.

08 Poza Transformerem

KAN, Mamba, mieszanka ekspertów, modele rozumujące i agenci. Czy skalowanie wciąż działa?

09 Szok DeepSeeka

Jak chiński model wstrząsnął rynkiem i jak potem wybuchły wojny o destylację.

10 Chiny dziś

GLM, Kimi i Seedance: wydajność, otwarte wagi i cena jako strategia.

11 Europa: na boisku czy na trybunach?

Mistral, Bielik, ASML i AI Act. Gdzie Europa prowadzi, a gdzie zostaje w tyle.

12 Mapa łańcucha wartości

Chipy, chmura, modele, aplikacje: gdzie są marże, gdzie czai się ryzyko i na co patrzeć.

Jeśli niektóre z tych słów wyglądają, jakby po klawiaturze przeszedł się kot, to dobrze – tak ma być. Każde wyjaśnimy w swoim czasie i żadne nie jest tak mądre, na jakie wygląda.

05

Jak czytać te rozdziały

Każdy rozdział opowiada od początku do końca jeden kawałek tej historii, razem z jej bohaterami. Można je czytać po kolei albo na wrywki, jak kto woli.

Wszystkie łączy kilka zasad:

- **Fakty są prawdziwe.** Sprawdziliśmy każdą datę, każde nazwisko i każdą liczbę, według stanu z października 2026 roku. Niepoważny jest wyłącznie ton.
- **Na zdjęciach są prawdziwi ludzie i prawdziwe maszyny**, a nie świecący na niebiesko mózg oplątany ścieżkami obwodu drukowanego.* Autorów zdjęć wymieniamy na końcu każdego rozdziału.
- **Co się dało narysować, to narysowaliśmy**, bo jeszcze nikt nie zrozumiał maszyny z samego akapitu.
- **Nie ma matematyki.** Jeśli już trzeba coś policzyć, to tylko tyle, ile zmieści się na serwetce.
- **W przypisach narrator mruczy pod nosem.** Wolno je pomijać. Kto jednak pomija drobny druk, ten potem podpisuje różne ciekawe rzeczy.
- **Każdy rozdział kończy się tak samo:** rachunkiem – kto płaci, kto zarabia i kogo chroni mur – a na koniec zostają trzy rzeczy do zapamiętania.

Zacznijmy od początku. Jest rok 1958, Buffalo w stanie Nowy Jork, a na scenie: psycholog, komputer wielkości pokoju i konferencja prasowa, która trochę się rozhułała.

* Istnieje prawo, do dziś nieuchylone, zgodnie z którym każdy artykuł o sztucznej inteligencji musi zdobić świecący niebieski mózg. Łamiemy je świadomie i jesteśmy gotowi ponieść konsekwencje.

CZĘŚĆ I: INTELIGENCJA

Jak uczą się maszyny

W którym maszynie wielkości pokoju obiecuje się złote góry, jedna książka psuje zabawę całemu pokoleniu, a uczenie się okazuje się schodzeniem z góry we mgle.

OKŁADKA: MGLISTA ŚCIEŻKA NA PILATUS W SZWAJCARII · GDMOLENA · CC BY-SA 4.0

01

Buffalo, 1958

W lipcu 1958 roku Marynarka Wojenna Stanów Zjednoczonych zwołała konferencję prasową, żeby pokazać światu maszynę, która odróżnia lewo od prawa.

Brzmi skromnie, ale robiło wrażenie.* Było to także dość osobliwe życzenie jak na marynarkę wojenną, skoro marynarze od paru stuleci mają na lewo i prawo własne, całkiem porządne słowa. Ale płaciła Marynarka, więc Marynarka dostała swoją konferencję.

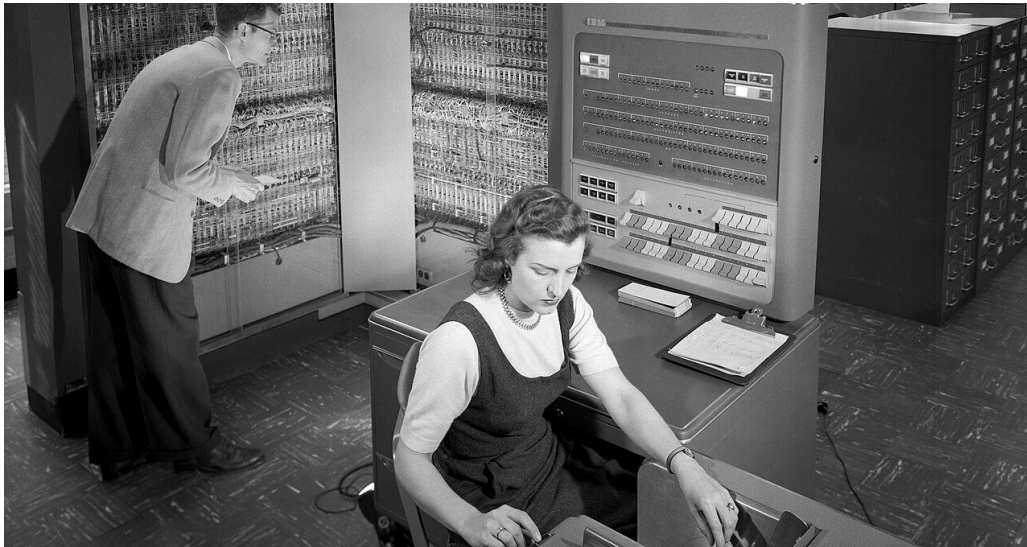
Maszyną był IBM 704: kilka ton komputera, który zajmował większą część pokoju i całą klimatyzację. Karmiono go stosem kart perforowanych. Jedne miały znak po lewej stronie, drugie po prawej. Nikt nie podał mu reguły, jak je rozróżnić. Po pięćdziesięciu próbach doszedł do tego sam.

Pokaz przygotował Frank Rosenblatt, psycholog z Cornell Aeronautical Laboratory w Buffalo w stanie Nowy Jork. Miał dwadzieścia dziewięć lat i jeszcze nie nauczył się ostrożności. Wynalazek nazwał Perceptronem – tak w filmach nazywa się maszyna, która w połowie seansu albo ratuje świat, albo zaczyna mu zagrażać. Wszyscy zgodnie mówią, że był genialny. Popełnił też klasyczny błąd geniuszy: myślał na głos przy dziennikarzach. Powiedział, że to pierwsza maszyna, która potrafi wpaść na oryginalny pomysł.

Resztę dopisała prasa, jak to prasa. „The New York Times” napisał, że według Marynarki potomkowie maszyny będą chodzić, mówić, widzieć, pisać, rozmnażać się

* Nawet dziś to więcej, niż można powiedzieć o niejednym kierowcy na wielopoziomowym parkingu.

i zdawać sobie sprawę z własnego istnienia.* Pod koniec roku „The New Yorker” uznał ją za pierwszego poważnego rywala ludzkiego mózgu.



IBM 704 na zdjęciu z 1957 roku, w rządowym laboratorium w Wirginii. Wszystko, co tu widać, to jeden komputer – z wyjątkiem dwojga ludzi, którzy są od karmienia. Przez cały dzień wykonywał mniej rachunków niż dzisiejszy wideodomofon w oczekiwaniu na listonosza.

FOT.: NASA • DOMENA PUBLICZNA

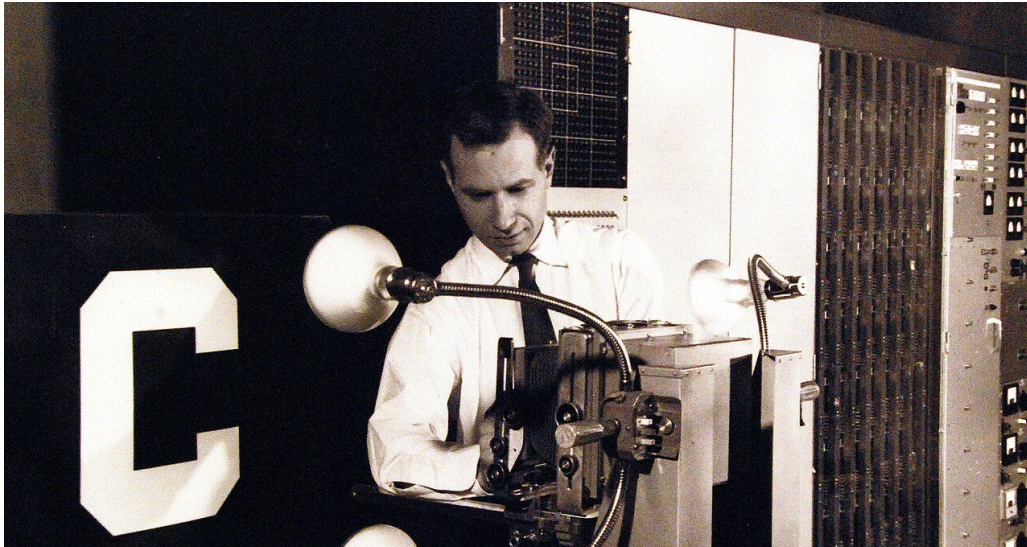
Odróżniała lewo od prawa.

Przepaść między tym, co maszyna umie w dniu pokazu, a tym, co według gazet będzie umiała do Świąt Bożego Narodzenia, warto dostrzec już teraz. Będzie za nami dreptać przez wszystkie dwanaście rozdziałów jak mały pies, który wciąż liczy na kiełbasę.

Dwa lata później pomysł dostał własne ciało. Mark I Perceptron miał zamiast oka kamerę z 400 czujników światła ułożonych w kwadrat dwadzieścia na dwadzieścia. Jego wiedza siedziała w rzędach pokręteł, a w trakcie nauki małe silniczki elektryczne naprawdę nimi kręciły – cierpliwy zwiedzający mógł stanąć w pokoju i słuchać, jak maszyna zmienia zdanie. Proszę zapamiętać te pokręta. Za chwilę staną się najważniejszymi przedmiotami w tej opowieści. Dziś są ich biliony, nikt już nie słyszy, jak się obracają, a cała sztuczka wciąż tkwi właśnie w nich.

Od maszyny z tego zdjęcia pochodzi niemal cała dzisiejsza sztuczna inteligencja: chatboty, generatory obrazów i rachunki za prąd. Zachwył przyszedł tylko jakieś sześćdziesiąt lat za wcześniej, a dziedzina słono za to zapłaciła. Zanim jednak dojdziemy do rachunku, spytajmy, co właściwie zbudował Rosenblatt. Do odpowiedzi wystarczy jeden pomysł, za to na nim stoi cała ta opowieść.

* To spore wymagania wobec kogokolwiek. Większość z nas w dobry dzień zalicza cztery punkty z sześciu, a z własnego istnienia zdaje sobie sprawę dopiero po kawie.



Mark I Perceptron ogląda literę C. Zdjęcie prasowe Marynarki Wojennej USA z czerwca 1960 roku. W podpisie Marynarka obiecywała maszynę, którą da się nauczyć rozpoznawać litery alfabetu. Miała rację. W końcu. Maszyna stoi dziś w Smithsonian, gdzie nikt już niczego od niej nie oczekuje.

FOT.: NATIONAL MUSEUM OF THE U.S. NAVY • DOMENA PUBLICZNA

02

Reguły i ich brak

Przez pięćdziesiąt lat oprogramowanie działało na jedną modłę. Człowiek spisywał reguły, komputer dostawał dane i wypływał odpowiedzi. Komputer jest w tym układzie bardzo szybkim i bardzo posłusznym urzędnikiem. Robi dokładnie to, co mu kazano, po wsze czasy – a ze szczególnym zapałem wtedy, gdy polecenie było błędne.*

Układ jest znakomity, o ile człowiek zna reguły. Kłopot w tym, że ludzie potrafią mnóstwo rzeczy, nie mając zielonego pojęcia, jak to robią. Własną matkę rozpoznasz bez trudu. A teraz spróbuj spisać reguły. „Dwoje oczu, a gdzieś pod nimi nos” opisuje ją całkiem trafnie – razem z większością ludzi, którzy nie są twoją matką, i sporą liczbą psów.

Uczenie maszynowe odwraca ten układ. Dajesz maszynie dane razem z poprawnymi odpowiedziami, a reguły maszyna wymyśla sama.

Weźmy przykład z naszego podwórka: komu dać kredyt. Po staremu regułę pisze analityk. Zatwierdzić, jeśli dochód przekracza określony próg, a klient nigdy nie przestał spłacać rat. Działa. Ale łatwo się sypie i nie widzi niczego, o czym analityk nie pomyślał – a że analitycy to ludzie, nie pomyśleli o większości rzeczy.

* Robi to przy tym kilka miliardów razy na sekundę. Tak oto drobna pomyłka w arkuszu kalkulacyjnym trafia na czołówki wieczornych wiadomości.

Dwa sposoby zagonienia komputera do roboty

PO STAREMU



Układ znakomity, o ile człowiek zna reguły.

PO NOWEMU



Też znakomity, o ile masz mnóstwo przykładów. Wyjaśnienie reguł płatne osobno.

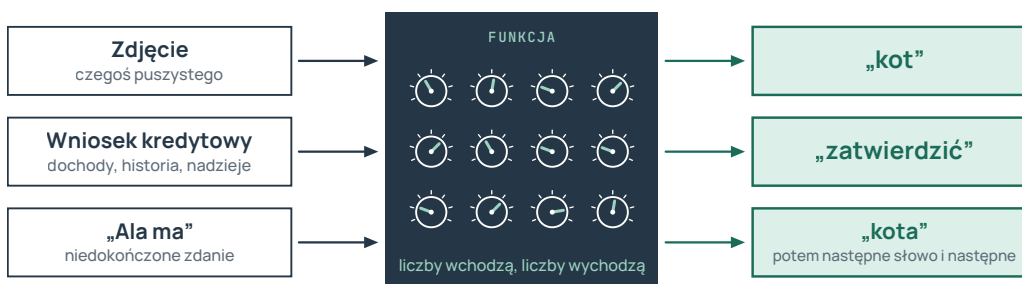
Cała rewolucja na jednej stronie. Klocki w obu rzędach są te same. Zielony przeszedł na drugą stronę.

Po nowemu pokazujesz modelowi miliony dawnych kredytów i to, jak każdy z nich się skończył. Model sam znajduje prawidłowości, po których poznać, że ktoś spłaci – także takie, jakich żaden człowiek by nie spisał. Ceną jest to, że znacznie trudniej wyjaśnić, dlaczego powiedział „tak”. Zapytaj analityka, a usłyszysz uzasadnienie. Zapytaj model, a dostaniesz w gruncie rzeczy wzruszenie ramion, wyliczone z dokładnością do jedenastu miejsc po przecinku. Proszę o tym pamiętać. Wróci w dalszej części opowieści, zwykle pod rękę z regulatorem.

Jeśli z tego rozdziału zapamiętasz tylko jedno, niech to będzie ta myśl. Jest mała i nie zajmie dużo miejsca.

Model to po prostu funkcja. Funkcja w sensie matematycznym to maszynka do mięsa: z jednej strony coś wkładasz, z drugiej wychodzi coś innego, a tego, co dzieje się w środku, lepiej nie oglądać przy obiedzie. Wkładasz zdjęcie, wychodzi „kot”. Wkładasz wniosek kredytowy, wychodzi „zatwierdzić”. ChatGPT i Claude to ten sam gatunek: wkładasz tekst, wychodzi następne słowo i tak w kółko, bardzo szybko – jak gość przy stole, który przeczytał cały internet i uparcie kończy za ciebie zdania. Niezwykłe jest w nich tylko to, jak ogromna jest ta funkcja i jak ją zbudowano.

Model to funkcja, a funkcja to pudełko z pokrętłami



Model bez obudowy. W środku nie ma nic poza pokrętłami. Cała sztuka polega na tym, jak je ustawić.

03

Jeden neuron, czyli członek komitetu

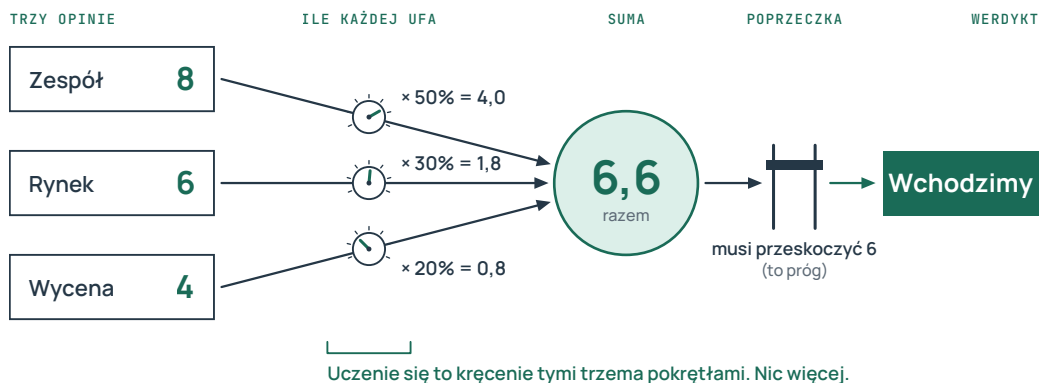
Co więc siedzi w środku? Podstawowy klocek to sztuczny neuron – i jest to rozczarowanie. Ludzie spodziewają się czegoś, co buczy, świeci, a może nawet ma zdanie o przyszłości ludzkości. Dostają dodawanie.

Neuron przyjmuje na wejściu kilka liczb. Każdą mnoży przez **wagę**, która określa, ile ta liczba ma do powiedzenia. Wszystko sumuje, dorzuca **próg** – można go traktować jak osobistą poprzeczkę – i rozstrzyga: odpalić albo nie odpalić. I tyle. Na odwrocie kartki nic więcej nie ma.

To stworzenie wszyscy znamy. Zasiada w komitetach inwestycyjnych.

Wyobraź sobie członka komitetu, który ocenia transakcję. Zespołowi daje osiem punktów, rynkowi sześć, a wycenie cztery. Ale nie ufa tym trzem rzeczom po równo. Zespół waży u niego połowę decyzji, rynek trzydzieści procent, a wycena dwadzieścia, bo z wycenami, jak wiadomo, bywa różnie.

Jeden neuron jako członek komitetu



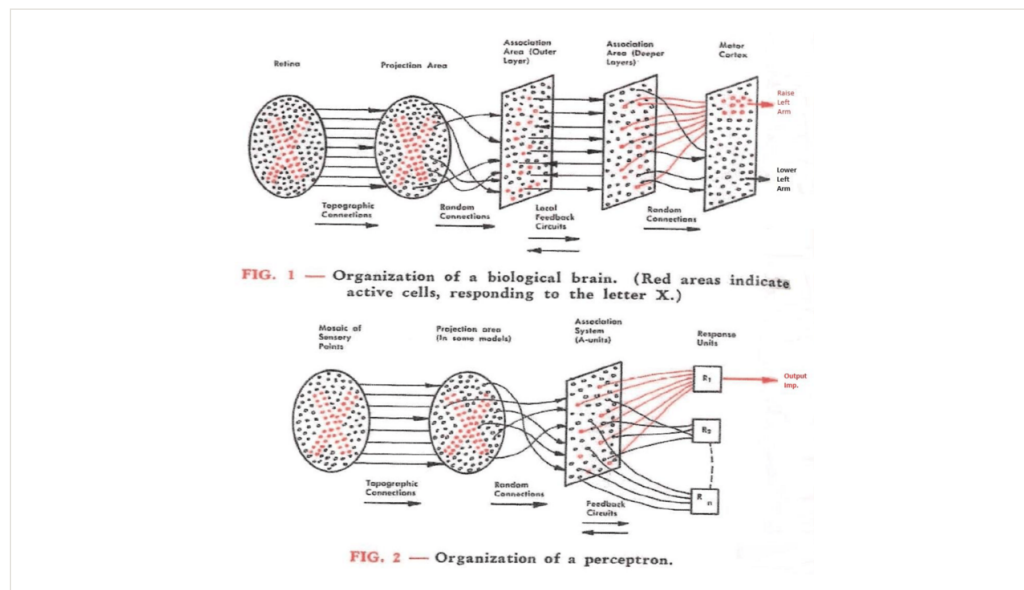
Sztuczny neuron podejmuje decyzję. Osiem razy połowa plus sześć razy trzy dziesiąte plus cztery razy jedna piąta daje 6,6. Poprzeczka wisi na szóstce. Wchodzimy.

Jego osobista poprzeczka to sześć. Suma wynosi 6,6. Werdykt: inwestujemy.

Oto cały neuron. A teraz najważniejsze: **uczenie się to regulowanie wag**. Kiedy nasz członek komitetu przyjrzy się, jak skończyły się dawne transakcje, może odkryć, że przeceniał zespół i nie doceniał wyceny – i odpowiednio przesunąć wagi.* Sieć neuro-nowa robi dokładnie to samo, tylko automatycznie, miliony razy, i nie trzeba jej przedtem zapraszać na lunch.

* Członkowie komitetów potrafią się przed tym bronić rękami i nogami. Neurony nie mają ani rąk, ani nóg, i głównie dlatego łatwiej je wytrenować.

Słowo „neuron” pożyczono od mózgu mniej więcej tak, jak żuraw budowlany pożyczyl nazwę od ptaka. Jakies podobienstwo rodzinne jest. Lepiej się na nim nie opierać.



Rysunek samego Rosenblatta z 1958 roku: u góry mózg, na dole jego perceptron. W tym podobieństwie był cały sens. Jak się później okazało – także cały problem.

FOT.: FRANK ROSENBLATT, 1958 • DOMENA PUBLICZNA

04

Komitet komitetów

Pojedynczy neuron zna dokładnie jedną sztuczkę. Przeciąga przez dane prostą linię i ogłasza: po tej stronie tak, po tamtej nie. To stworzenie z linijką i jednym niezachwianym poglądem.

Co to znaczy, pokaże mały eksperyment. Można go przeprowadzić na papierze – tam trudniej coś zepsuć.

Wyobraź sobie kartkę usianą kropkami. Pośrodku skupisko pomarańczowych, wokół pierścieni niebieskich – jak jajko sadzone narysowane przez kogoś, komu je tylko kiedyś opisano. Trzeba oddzielić pomarańczowe od niebieskich. Daj to zadanie jednemu neuronowi, a narysuje swoją jedyną prostą. Gdziekolwiek ją poprowadzi, niebieskie zostaną po obu stronach. Jedną kreską od linijki żółtka z jajka nie wykroisz. Neuron i tak próbuje, jak długo mu pozwolisz, ze spokojnym uporem człowieka, który pcha drzwi z napisem CIĄGNAĆ – a jego błąd nie maleje ani o włos.

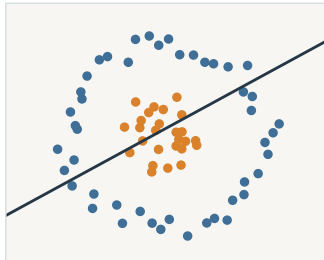
Teraz daj to samo zadanie czterem neuronom naraz. Każdy rysuje swoją linię. Cztery linie tworzą ramkę, ramka otacza żółtko i po kilku sekundach błąd leci na łeb na szyję.

A teraz coś trudniejszego: dwa kolory zwinięte w spiralę, jak drożdżowy ślimak z cynamonem. Cztery neurony nic tu nie wskórają. Potrzeba trzech warstw po osiem

neuronów, mniej więcej 180 pokręteł i minuty czy dwóch cierpliwości. I nawet wtedy nie każda próba się udaje.

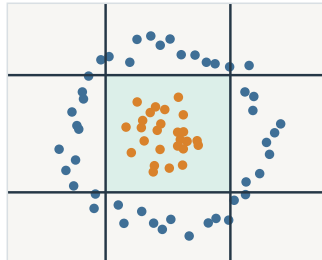
Problem jajka sadzonego

JEDEN NEURON: JEDNA LINIA



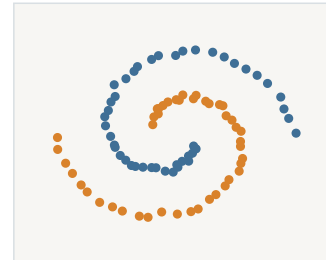
Gdzie nie przyłożyć linijki, niebieskie są po obu stronach. Błąd ani drgnie.

CZTERY NEURONY: CZTERY LINIE



Cztery linie tworzą ramkę, a ramka otacza żółtko. Kilka sekund i gotowe.

SPIRALA, CZYLI ŚLIMAK



Trzy warstwy po osiem neuronów, minuta cierpliwości i trochę szczęścia.

Śniadanie na trzy sposoby. Jedną linią żółtka nie ogrodzisz. Czterema – owszem. Na ślimaka trzeba już komitetu komitetów.

W tym małym eksperymencie mieści się cała branża. **Pojemność to liczba pokręteł. Koszt to czas potrzebny, żeby je ustawić.** Wszystko, co nastąpiło później, to ten sam eksperyment, tylko z coraz większym jajkiem.

Przełom polega więc na tym, żeby ułożyć neurony w warstwy, z których każda karmi następną: zbudować komitet komitetów. Do oryginalnych dokumentów zaglądają tylko najmłodszy stażem, na samym dole, a każde piętro przekazuje wyżej streszczenie. Każdy, kto pracował w dużej firmie, zna ten układ. Zaskakuje tylko jedno: tutaj to działa.

Pokaż takiej sieci miliony zdjęć, a stanie się coś niezwykłego. Pierwsza warstwa nauczy się wypatrywać krawędzi. Następna złoży krawędzie w kształty. Ostatnia rozpozna całe przedmioty i oznajmi, że to kot. Nikt jej nie powiedział, co to krawędź. Sama znalazła krawędzie, bo z nimi częściej trafiała – trochę tak, jak nikt nie tłumaczy młodszemu analitykowi, po czym poznać, że w transakcji coś śmierdzi, a on po trzech latach w firmie czuje to przez zamknięte drzwi.

Tylko tyle znaczy słowo „głębokie” w nazwie „uczenie głębokie”. Nie chodzi o głębię myśli. Chodzi o to, że warstw jest dużo – co rozczerowało niejednego filozofa i co najmniej jeden dział marketingu.

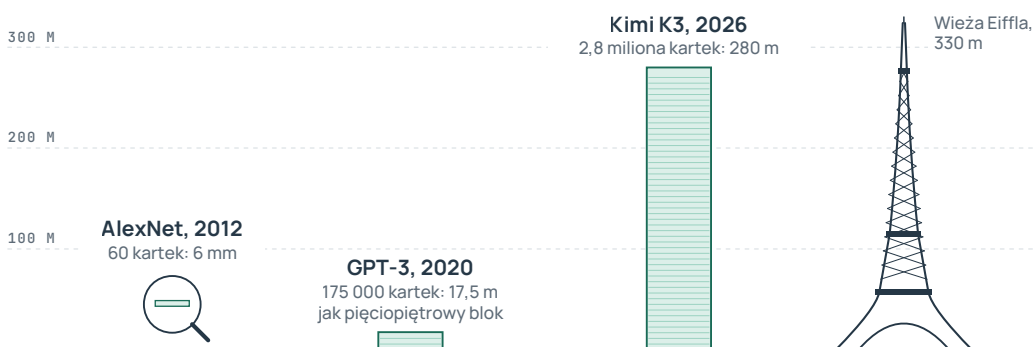
Każda waga w sieci to jedno pokrętko, a branża nazywa te pokrętko **parametrami**, zapewne dlatego, że „pokrętko” nie brzmiały dość drogo. Ich liczba trochę wymknęła się spod kontroli:

- AlexNet – wytrenowany w 2012 roku na tamtych dwóch kartach do gier – miał ich około 60 milionów.
- GPT-3 w 2020 roku miał ich 175 miliardów.

- Kimi K3, wypuszczony w lipcu 2026 roku przez chińską firmę Moonshot AI, ma ich według producenta 2,8 biliona.

To mniej więcej 47 000 razy więcej niż AlexNet, i to w czternaście lat. Liczby tej wielkości ześlizgują się z wyobraźni jak jajko sadzone z talerza, sięgnijmy więc po artykuły biurowe. Przyjmijmy jedną kartkę zwykłego papieru do drukarki na każdy milion pokręteł i ułożmy kartki w stos.

Jedna kartka papieru na każdy milion pokręteł



AlexNet to cieniutka broszurka. GPT-3 to stos wysoki jak pięciopiętrowy blok. Kimi K3 zaprowadziłby nas prawie na szczyt wieży Eiffla, a z góry widać głównie rachunek za prąd.

Każde z tych pokręteł trzeba ustawić w trakcie treningu i dlatego opowieść o arytmetyce prędzej czy później zamienia się w opowieść o pieniądzach i prądzie.*

05

Mgła w górach

Jak ustawić blisko trzy biliony pokręteł? Nie ręcznie i już nie silniczkami.† Potrzebne są trzy pomysły i to one są sednem sprawy.

Po pierwsze, zmierz, jak bardzo się mylisz. Model ocenia na siedemdziesiąt procent szansę, że kredyt nie zostanie spłacony. Tymczasem klient spłacił go co do grosza. Rozbieżność wyraża się jedną liczbą, zwaną **stratą**. Duża strata: pudło. Mała strata: prawie w dziesiątkę. Trening ma w życiu dokładnie jedną ambicję – zbić tę liczbę na milionach przykładów. Wyobraź sobie synoptyka, który po każdej prognozie dostaje ocenę z dokładnością do kilku miejsc po przecinku i nie wolno mu wyjść z pokoju, dopóki się nie poprawi. Niezbyt to miłe. Ale właśnie dlatego metoda działa.

* Mała przestroga: liczba parametrów mówi, jak duży jest silnik i ile pali. Nie mówi, jak dobrze ktoś tym jeździ.

† Przy jednym pokrętle na sekundę, bez przerw na herbatę, silniczki potrzebowałyby około 89 000 lat, żeby obejść wszystkie choćby raz. Trening obchodzi je wiele, wiele razy.

Po drugie, idź w dół. Wyobraź sobie, że stoisz na górze w gęstej mgle i wolałbyś być w dolinie, gdzie strata jest mała i gdzie, jak należy sądzić, stoi schronisko. Okolicy nie widzisz. Możesz tylko wyczuć pod butami, w którą stronę opada teren, i zrobić jeden krok w dół. Potem znowu. I znowu, kilka milionów razy. Nazywa się to **spadkiem gradientu** i jest zajęciem dużo mniej sportowym, niż sugeruje nazwa.



Spadek gradientu w wersji alaskańskiej. Doliny nie widać. Ale czuć, gdzie jest dół.

FOT.: JUSTIN MORGAN • CC BY-SA 2.0

Wielkość kroku ma ogromne znaczenie. Za duży – i przeskoczysz dolinę i wylądujesz na przeciwnym zboczu. W naszym eksperymencie ze spiralą wystarczy powiększyć krok stokrotnie, a błąd tylko podskakuje w górę i w dół, i nigdy się nie uspokaja. Za mały – i za rok o tej porze wciąż będziesz w drodze. Dobór kroku to osobne rzemiosło, a kto je opanował, temu odpowiednio płacą.

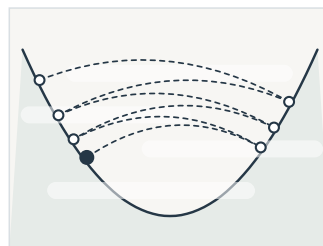
Trzy sposoby schodzenia z góry we mgle

KROK ZA MAŁY



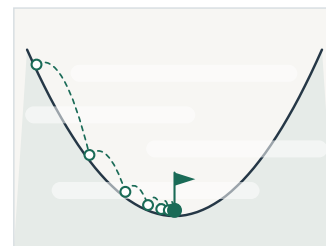
Bezpiecznie, tylko że za rok o tej porze nadal w drodze.

KROK ZA DUŻY



Przesadza dolinę jednym susem i ląduje na drugim zboczu. I od nowa.

W SAM RAZ



Rozsądny krok, z każdym razem nieco krótszy. Kolacja w schronisku.

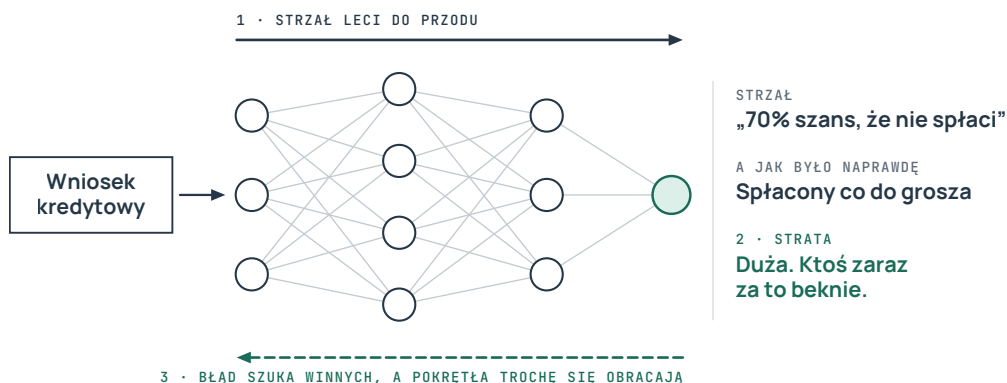
Trzy sposoby schodzenia z góry, której nie widać. Branża nazywa długość kroku współczynnikiem uczenia i zarwała przez niego niejedną noc.

Po trzecie, znajdź winnych. Skąd przy bilionach pokręteł wiadomo, którymi z nich pokręcić? Odpowiedź nazywa się **propagacją wsteczną**, a rozstawili ją w 1986 roku

David Rumelhart, Geoffrey Hinton i Ronald Williams. Błąd z wyjścia wędruje przez sieć wstecz, warstwa po warstwie, a każde pokrętło dowiadyuje się, ile dołożyło do pomyłki – i dostaje swoją część poprawki.

Szukanie winnych to w końcu nasz sport narodowy, tyle że tu wreszcie do czegoś służy. Pomyśl o odprowie po nieudanej transakcji. Przechodzisz cały proces od końca, ustalasz, kto ile dołożył, i każdy trochę się poprawia. Różnica polega na tym, że w sieci neuronowej nikt się nie obraża, nikt nie zwala wszystkiego na rynek, nikt nie twierdzi, że najważniejszy wniosek dotyczy komunikacji, a całe zebranie trwa ułamek sekundy.

Szukanie winnych, czyli jak błąd wraca do pokręteł



Propagacja wsteczna na jednym obrazku. Strzał idzie w jedną stronę, błąd wraca w drugą. Każde pokrętło dostaje po uszach dokładnie tyle, ile mu się należy, i obraca się odrobinę. Potem wszystko od nowa, kilka miliardów razy.



David Rumelhart na konferencji o sieciach neuronowych w Seattle w lipcu 1991 roku. Pięć lat wcześniej wraz z Hintonem i Williamsem pokazał światu sprawny sposób szukania winnych.

FOT.: ROLF KICKUTH • CC BY-SA 4.0

Warto zobaczyć, jak to szukanie winnych wygląda od środka, bo sztuczka jest prostsza, niż głosi jej sława. Weźmy dowolne pokrętło. Musi się dowiedzieć jednego: gdyby przekręcić je o włos w prawo, czy strata by wzrosła, czy spadła, i o ile? To właśnie nachylenie zbocza pod butami z naszej zamglonej góry, tyle że dla jednego pokrętła. Dałoby się je zmierzyć doświadczalnie: trącić pokrętło, przeliczyć całą sieć jeszcze raz i zobaczyć, co wyjdzie. Przy trzech bilionach pokręteł oznaczałoby to trzy

biliony przeliczeń na każdy krok w dół zbrocza. Przy czymś takim nawet silniczek Rosenblatta wyglądają na żwawe.

Propagacja wsteczna dostaje wszystkie odpowiedzi z jednej podróży do tyłu, a udaje się jej to dlatego, że nikt nie musi wiedzieć więcej niż to, co leży na jego biurku. Weźmy najmniejszy komitet, jaki da się zwołać: młodszego analityka, który czyta wniosek i wystawia mu ocenę, starszą partnerkę, która waży tę ocenę własnym pokrętkiem, i drugie pokrętło, między nią a ostatecznym werdyktem. Werdykt brzmi: siedemdziesiąt procent szans, że kredyt nie zostanie spłacony. Kredyt spłacono. Werdykt przeszedł o 0,7 i tu zaczyna się szukanie winnych.

Pierwsze idzie pokrętło najbliższe werdyktu. To, ile dołożyło do pomyłki, zależy od dwóch rzeczy: o ile przeszedł werdykt i ile przez to pokrętło przeszedł. Pokrętło, przez które szła głośna opinia, mocno wypchnęło werdykt w górę. Takie, przez które szedł szept, prawie wcale. Jego część to więc jedno pomnożone przez drugie. Potem partnerka dowiadyuje się, o ile sama się pomyliła: o przeszerzenie pomnożone przez pokrętło, które łączy ją z werdyktem. Jeśli jej zdanie liczyło się w połowie, połowa przeszerzenia idzie na jej konto. Teraz ona robi dokładnie to samo z pokrętkiem pod sobą i podaje niżej, młodszemu analitykowi, mniejszą karteczkę, a on robi to samo ze swoimi pokrętkami. Nikt nie patrzy dalej niż jedno biurko w górę i jedno w dół – w większości firm nazywa się to pracą w silosach, a tutaj wydajnością. Karteczka z liczbą po prostu wędruje przez budynek z góry na dół, przy każdym biurku przemnożona, aż każde pokrętło pozna swoją liczbę.

Kto płaci za przeszerzenie

1 · W PRZÓD: KAŻDA OPINIA RAZY POKRĘTKO



3 · KAŻDE POKRĘTKO OBRACA SIĘ POD PRĄD SWOJEJ CZĘŚCI, O JEDNĄ DZIESIĄTĄ

Pokrętła przechodzą z 0,7 i 0,5 na 0,63 i 0,40. Następnym razem werdykt to 51 procent, nie 70. Nikt nie musiał w tym celu przeliczać sieci od nowa.

Najmniejszy komitet świata po nieudanych kredycie. W przód idą opinie, każda pomnożona przez pokrętło. Wstecz wraca przeszerzenie, przemnażane przy każdym biurku, i każde pokrętło poznaje swoją część w jednej podróży.

Potem każde pokrętło obraca się odrobinę pod prąd swojej części i następny strzał jest trochę celniejszy. Na rysunku wystarczy obrócić każde pokrętło o jedną dziesiątą jego części, żeby werdykt spadł w jednym kroku z 70 do mniej więcej 51 procent. Prawdziwe sieci robią kroki dużo mniejsze i dużo liczniejsze.

Matematycy nazywają to regułą łańcuchową, a w szkole uczą jej jako pochodnej funkcji złożonej, zwykle nie wspominając, że jest warta bilion dolarów. Wynikają z niej trzy rzeczy. Po pierwsze, neuron, który przeskakuje między „tak” a „nie”, jak nasz członek komitetu z początku rozdziału, byłby tu bezużyteczny: przekręcenie pokrętła o włos niczego by nie zmieniło i błąd nie miałby którędy wrócić. Dlatego neurony w trenowanej sieci nie przeskakują, tylko płynnie przechodzą od „raczej nie” do „raczej tak”, a ich zdanie da się odrobinę przesunąć. Po drugie, karteczka jest mnożona przy każdym biurku, więc w wysokim budynku może po drodze w dół skurczyć się prawie do zera. Rozdział 3 spotka się z tym kłopotem na szkolnym podwórku. Po trzecie, droga wstecz wymaga mniej więcej dwa razy więcej liczenia niż droga w przód, więc jeden krok treningu kosztuje mniej więcej trzy razy tyle co zwykła odpowiedź – i to zanim ktokolwiek zacznie liczyć kroki.*

Na koniec jeszcze jedna obserwacja z tego samego eksperymentu. Wróćmy do jajka sadzonego i samotnego neuronu, który sobie z nim nie poradził. Powiedzmy mu teraz o każdej kropce jedną rzecz więcej: jak daleko leży od środka kartki. Neuron od razu rozprawia się z jajkiem – sam, bez żadnego komitetu. Dobre dane biją brutalną siłę. Proszę schować tę myśl w bezpiecznym miejscu. W następnym rozdziale zamieni się w pieniądze.

06

Wielkie pomysły, nie ta dekada

Oto zagadka. Wszystko, o czym była dotąd mowa – neurony, warstwy, schodzenie z góry i szukanie winnych – istniało już w 1986 roku. Boom na sztuczną inteligencję zaczął się w 2012. Co wszyscy robili przez te ćwierć wieku?

Głównie marzli.

Branża mówi o **zimach sztucznej inteligencji**: długich okresach, gdy pieniądze odpływały, konferencje się kurczyły, a kto przy kolacji przyznał się do pracy nad sieciami neuronowymi, ten zbierał taką samą współczującą ciszę jak ktoś, kto wyznaje, że pisze powieść. Były dwie takie zimy.

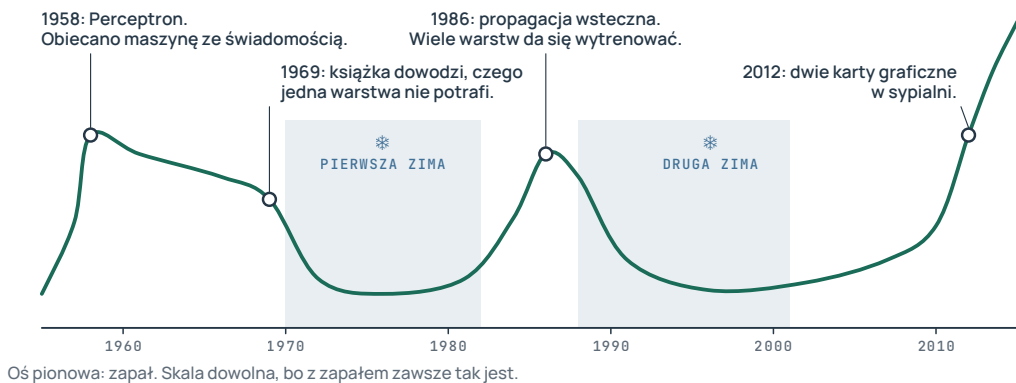
Pierwszy przymrozek przyszedł z książki. W 1969 roku Marvin Minsky i Seymour Papert, obaj z MIT, wydali książkę *Perceptrons*, w której z wielką matematyczną starannością dowiedli, że pojedyncza warstwa neuronów nie nauczy się nawet tak prostej zasady jak „albo jedno, albo drugie, ale nie oba naraz”.† Mieli rację. To znowu

* Jak większość dobrych pomysłów, i ten wymyślono kilka razy, zanim stał się sławny. Fiński student Seppo Linnainmaa opisał ogólną metodę w pracy magisterskiej w 1970 roku, a Amerykanin Paul Werbos podsunął w doktoracie z 1974 roku, żeby zastosować ją do sieci neuronowych. Artykuł z 1986 roku dorzucił dowód, że to działa, i to, że środkowe warstwy uczą się przy tym pożytecznych rzeczy, których nikt im nie podpowiedział. A to akurat da się sprzedać.

† Zupa albo sałatka, jak stoi w karcie dań. Każdy kelner tego dopilnuje. Pojedynczy neuron – nie.

nasze jajko sadzone: jedna prosta linia nie wystarczy. Że lekarstwem są kolejne warstwy, podejrzewano już wtedy, ale nikt jeszcze nie wiedział, jak takie warstwy trenować, a pieniądze nie czekały, aż ktoś się dowie. Pieniądze rzadko czekają.

Zapał do sieci neuronowych, 1955–2015



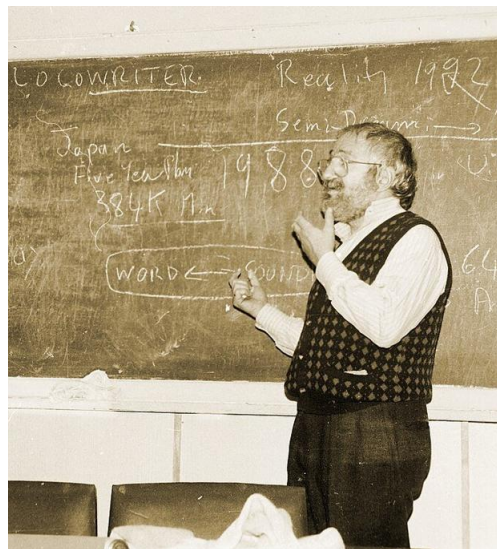
Sześćdziesiąt lat pogody. Dwa razy lato, dwa razy zima, a potem to, co mamy teraz, czymkolwiek to jest.

Rosenblatt nie doczekał odwilży. Zginął w 1971 roku w wypadku na żagłówce, w dniu swoich czterdziestych trzech urodzin. Wyśmiewano go za to, że obiecuje po swoim pomysły za dużo. Niemal każda maszyna w tej opowieści działa właśnie na nim.



Marvin Minsky w 2012 roku. On i Rosenblatt chodzili do tego samego liceum na Bronksie. Świat nauki to małe miasteczko.

FOT.: JOI ITO • CC BY 2.0



Seymour Papert podczas wykładu w Moskwie w 1987 roku. Był już wtedy bardziej znany z tego, że uczył dzieci programowania, niż z tego, że wsadził całą dziedzinę nauki do zamrażarki.

FOT.: SHEN-MONTELLIER • CC BY-SA 4.0

Odwilż przyszła w 1986 roku i nazywała się propagacja wsteczna: nareszcie było wiadomo, jak trenować wiele warstw. Wrócił zapał, a z nim pieniądze, które przezimowały gdzieś w ciepłych krajach. Długo to nie trwało. W latach dziewięćdziesiątych

pomysły pięknie działały na szkolnych przykładach i wykladały się na prawdziwych – co zna każdy, kto przesiedział choć jeden pokaz nowego produktu – i nastąpiła druga zima.

Czego więc brakowało w 1986 roku? Nie teorii. Trzech składników:

- **Danych.** Internet nie wyprodukował jeszcze miliardów zdjęć, zdań i opisów, na których można by się uczyć.
- **Mocy obliczeniowej.** Nikt jeszcze nie zauważył, że karty graficzne, budowane z myślą o grach wideo, znakomicie radzą sobie z jedyną rzeczą, której potrzebują sieci neuronowe: z mnożeniem olbrzymich tabel liczb.
- **Sztuczek.** Nikt jeszcze nie wpadł na garść mało efektywnych technik treningowych, dzięki którym głębokie sieci się nie wykladają.

Dziedzina miała przepis, tylko nie miała z czego ugotować. W 2012 roku wszystkie trzy składniki pojawiły się naraz. Wielkie pomysły, nie ta dekada. A potem, całkiem nagle, ta właściwa.

07

Rachunek

Zmiermy teraz kapelusz i pójdźmy za pieniędzmi.

Oto coś, co zwykle zaskakuje. Wytrenowany model jest dosłownie **plikiem z liczbami**: ustawieniami wszystkich pokręteł, zapisanymi po kolei. Plik Kimi K3 jest bardzo duży, ale to wciąż tylko plik. Można by go postawić na półce. Można by go nawet zgrać na dysk tak mały, że łatwo go zgubić między poduszkami kanapy.*

Gdzie zatem jest wartość? W tych liczbach – i w danych oraz mocy obliczeniowej, które spalono, żeby je uzyskać. Stąd rozróżnienie, które będzie nam towarzyszyć do końca opowieści:

- **Koszt treningu** płaci się raz, z góry, i jest ogromny. To koszt ustawienia wszystkich pokręteł.
- **Koszt inferencji** płaci się za każdym razem, gdy model odpowiada na pytanie. Rośnie razem z liczbą pytań, po wsze czasy.

* Przy dwóch bajtach na pokrętko wychodzi około 5,6 terabajta. Kanapy połykały już większe rzeczy.

Dwa rodzaje rachunków

TRENING: PŁACONY RAZ



INFERENCJA: PŁACONA, ILEKROĆ KTOKOLWIEK O COKOLWIEK ZAPYTA



Każdy z osobna jest drobny. Przychodzą za każdym razem, gdy model na coś odpowie, i nigdy nie przestają przychodzić.

CZAS, UŻYCIE I CAŁY DAŁSZY ŻYWIOT FIRMY

Jeden rachunek przychodzi raz i łąduje w ramce na ścianie. Drugi przychodzi, ilekroć ktokolwiek zapyta model o cokolwiek, i do ramki nie trafia.

Różne firmy zarabiają po różnych stronach tej linii, a w niejednej prezentacji dla inwestorów bardzo trudno się doczytać, po której stronie stoi sama firma.

I pytanie do przemyslenia w domu. Skoro wartością modelu jest plik, co się stanie, kiedy ktoś rozda ten plik za darmo? Albo go skopiuje? Odłóżmy to na razie. W rozdziałach dziewiątym i dziesiątym z tej jednej myśli zrobi się cała fabuła.

08

Trzy rzeczy do zapamiętania

- 01 **Sieć to funkcja z pokrętłami.** Dane wchodzą, odpowiedź wychodzi, a pomiędzy nimi siedzi mnóstwo liczb do ustawienia.
- 02 **Uczenie się to kręcenie pokrętłami, aż błąd zmaleje.** Zmierz pomyłkę, wycuj nachylenie, zrób krok w dół, powtórz.
- 03 **Rewolucja wzięła się z danych i sprzętu, a nie z nowego pomysłu.** Pomysły leżały na półce od 1986 roku i czekały, aż świat do nich dorośnie.

Występują

KTO	CO ZROBIŁ	CO BYŁO DALEJ
Frank Rosenblatt	Psycholog w Cornell Aeronautical Laboratory w Buffalo. W 1958 roku pokazał światu Perceptron.	Później wykładał na Uniwersytecie Cornella. Zmarł w 1971 roku w wieku 43 lat.
Mark I Perceptron	Pomysł, który stał się szafą: 400 czujników światła, 512 jednostek ukrytych i wagi zapisane na pokrętkach, którymi kręciły silniczek. Pierwszy publiczny pokaz w czerwcu 1960 roku.	Przeszedł na emeryturę do Narodowego Muzeum Historii Amerykańskiej, należącego do Smithsonian.
Marvin Minsky	Współzałożyciel laboratorium sztucznej inteligencji w MIT i współautor Perceptrons (1969).	Zmarł w 2016 roku w wieku 88 lat.
Seymour Papert	Współautor Perceptrons, później sławny dzięki językowi Logo, w którym dzieci uczyły się programować.	Zmarł w 2016 roku.
David Rumelhart, Geoffrey Hinton, Ronald Williams	Autorzy artykułu z 1986 roku, który rozstawił propagację wsteczną.	Hinton został współlaureatem Nagrody Nobla z fizyki za 2024 rok i wraca w następnym rozdziale.

KONIEC PRÓBKII

Tu kończy się próbka. Opowieść idzie dalej.

Za Wami wstęp i pierwszy rozdział, mniej więcej jedna trzynasta książki. Pozostałe jedenaście rozdziałów prowadzi od kart graficznych przez elektrownie i wideo aż do pytania, kto na tym wszystkim zarabia.

CZĘŚĆ I: INTELIGENCJA Jak działa sztuczna inteligencja

- 02 Maszyny, które widzą**
Sieci konwolucyjne i przełomowy rok 2012 w konkursie ImageNet, kiedy karty do gier weszły do gry o sztuczną inteligencję.
- 03 Attention Is All You Need**
Od sieci rekurencyjnych do Transformera, architektury skrojonej na miarę sprzętu, który liczy równolegle.
- 04 Jak zbudować model językowy**
Pretrening, prawa skalowania i RLHF. Czego naprawdę trzeba, żeby model wytrenować, a czego, żeby potem utrzymać go w ruchu.

CZĘŚĆ II: MOC OBLICZENIOWA I ENERGIA Warstwa fizyczna

- 05 Ekonomia krzemu**
Dlaczego GPU rządzą, dlaczego tyle kosztują i kto trzyma w rękę łańcuch dostaw.
- 06 Prąd i obliczenia alternatywne**
Ile prądu zjada sztuczna inteligencja i kto chce to zmienić: układy analogowe, fotoniczne i neuromorficzne.

CZĘŚĆ III: NA PIERWSZEJ LINII Generatywna sztuczna inteligencja i to, co po niej

- 07 Obrazy w ruchu**
GAN-y, dyfuzja i Stable Diffusion, a także to, dlaczego wideo to jak dotąd najtwardszy orzech do zgryzienia.
- 08 Poza Transformerem**
KAN, Mamba, mieszanka ekspertów, modele rozumujące i agenci. Czy skalowanie wciąż działa?

CZĘŚĆ IV: WYŚCIG O SUWERENNOŚĆ USA, Chiny, Europa

- 09 Szok DeepSeeka**
Jak chiński model wstrząsnął rynkiem i jak potem wybuchły wojny o destylację.
- 10 Chiny dziś**
GLM, Kimi i Seedance: wydajność, otwarte wagi i cena jako strategia.
- 11 Europa: na boisku czy na trybunach?**
Mistral, Bielik, ASML i AI Act. Gdzie Europa prowadzi, a gdzie zostaje w tyle.
- 12 Mapa łańcucha wartości**
Chipy, chmura, modele, aplikacje: gdzie są marże, gdzie czai się ryzyko i na co patrzeć.

Całą książkę kupicie na neuron.bazyl.net

340 stron, PDF i EPUB, po polsku i po angielsku, bez DRM. Cena: **29 zł**. Płatność BLIK-iem, przelewem lub kartą; plik do pobrania od razu po zapłacie.

Tę próbkę wolno swobodnie przesyłać dalej i udostępniać, byle w niezmienionej postaci.

KONIEC PRÓBKII

Ilustracje w tej próbce

Fotografie i reprodukcje są dziełami innych autorów i pozostają na swoich licencjach. Numery stron podane na stronie redakcyjnej odnoszą się do pełnego wydania; dla zdjęć z tej próbki autora, źródło i licencję podajemy poniżej. Zmiany: skalowanie i kadrowanie, a w okładkach rozdziałów i na stronie tytułowej także przyciemnienie i tekst nałożony na zdjęcie. Rysunki i wykresy: autor.

Strona tytułowa: krzem karty GeForce GTX 580. Fritzchens Fritz, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:NVIDIA@40nm@Fermi@GF110@GeForce_GTX_580@UA10B338_1041A1_N2Y540.000_GF110-375-A1_DSC04152_%2827576804994%29.jpg. CC0 (<https://creativecommons.org/publicdomain/zero/1.0/>).

Okładka wprowadzenia: centra danych pod Ashburn w Wirginii. Theodore Christopher, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Data_centers_in_Ashburn.jpg. CC0 (<https://creativecommons.org/publicdomain/zero/1.0/>).

Okładka rozdziału 1: mglista ścieżka na Pilatus w Szwajcarii. Gdmolena, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Foggy_path_from_Alpnach_to_Pilatus_mountain.jpg. CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>).

GeForce GTX 580. TheStriker, Wikimedia Commons, <https://commons.wikimedia.org/wiki/File:EVGAGeforceGTX580.jpg>. CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>).

Centrum danych Google'a, Council Bluffs, Iowa. Chad Davis, Wikimedia Commons, [https://commons.wikimedia.org/wiki/File:Google_Data_Center,_Council_Bluffs_Iowa_\(49062863796\).jpg](https://commons.wikimedia.org/wiki/File:Google_Data_Center,_Council_Bluffs_Iowa_(49062863796).jpg). CC BY 2.0 (<https://creativecommons.org/licenses/by/2.0/>).

IBM 704, 1957. NASA, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:IBM_Electronic_Data_Processing_Machine_-_GPN-2000-001881.jpg. domena publiczna.

Mark I Perceptron. National Museum of the U.S. Navy, Wikimedia Commons, [https://commons.wikimedia.org/wiki/File:330-PSA-80-60_\(USN_710739\)_\(20897323365\).jpg](https://commons.wikimedia.org/wiki/File:330-PSA-80-60_(USN_710739)_(20897323365).jpg). domena publiczna.

Schematy mózgu i perceptronu. Frank Rosenblatt, 1958, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Organization_of_a_biological_brain_and_a_perceptron.png. domena publiczna.

Turysta we mgle, Alaska. Justin Morgan, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Flat_Top_Mountain_Hike.jpg. CC BY-SA 2.0 (<https://creativecommons.org/licenses/by-sa/2.0/>).

David Rumelhart, 1991. Rolf Kickuth, Wikimedia Commons, <https://commons.wikimedia.org/wiki/File:DavidRumelhart-IJCNNseattle1991-07-08.jpg>. CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>).

Marvin Minsky, 2012. Joi Ito, Wikimedia Commons, [https://commons.wikimedia.org/wiki/File:Marvin_Minsky_\(1\).jpg](https://commons.wikimedia.org/wiki/File:Marvin_Minsky_(1).jpg). CC BY 2.0 (<https://creativecommons.org/licenses/by/2.0/>).

Seymour Papert, 1987. Shen-montpellier, Wikimedia Commons, <https://commons.wikimedia.org/wiki/File:Papert-moscow-1987.jpg>. CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>).
