

Inżynieria Mechatroniczna

Podstawy Sztucznej Inteligencji i Uczenia Głębokiego:

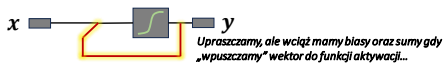
12: Transformers

Ziemowit Dworakowski
AGH w Krakowie

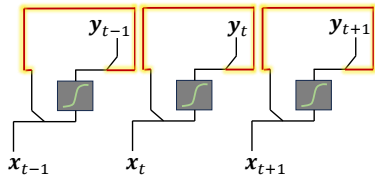
1

Sieci rekurencyjne – ponownie

Rozważmy bardzo prostą sieć rekurencyjną (wyłącznie jedna **warstwa**)

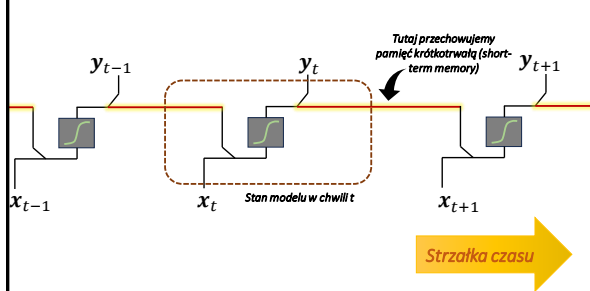


I przerysujmy nasz schemat w dość dziwny sposób:

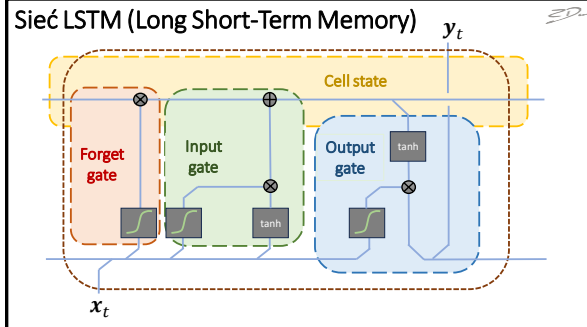


2

Sieci rekurencyjne – ponownie



3



4

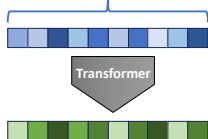
LSTMy są szybkie i skuteczne w zadaniach predykcji sekwencji i mogą zachowywać przez długi czas wybrane informacje, ale:

- Wymagają nauki sekwencyjnej (czasochłonne!)
- Nie uczą się łatwo złożonych zależności
- Nie skalują się zbyt łatwo

5

Czym jest transformer?

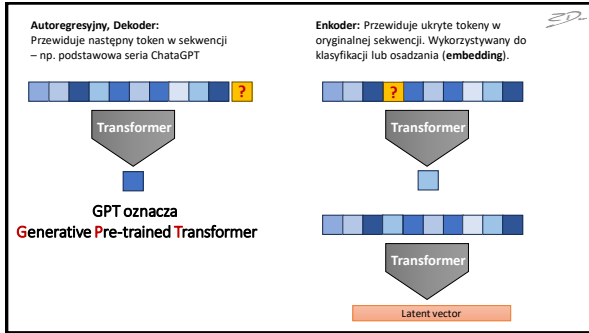
Używa uwagi (**attention**) to by przetwarzać całą sekwencję wejścia



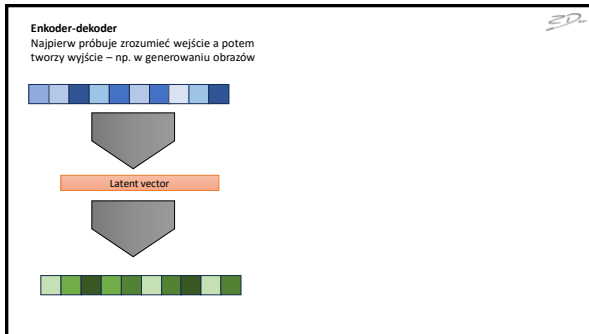
System, który transformuje jedną sekwencję w inną, na przykład:

Text-to-text,
text-to-speech,
text-to-image,
speech-to-text,
...
UnitData-to-situation
(AlphaStar w grze Starcraft 2)
...
etc.

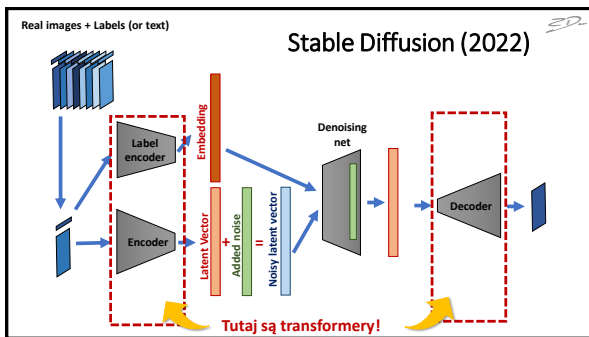
6



7



8



9

Więc czym jest uwaga (**attention**)?

He smiled and accepted the award, knowing it was meant for someone else.

Tekst rozkładany jest na „tokens” (dłuższe słowa na kilka, ale dla uproszczenia przyjmijmy, że token = słowo).
Sekwencja tokenów jest ważna
(każdy ma przypisaną pozycję w tekście)

10

Więc czym jest uwaga (**attention**)?

He smiled and accepted the award, knowing it was meant for someone else.

Przejsięcie 1: Które słowa „aktualizują” pierwszy token? (powtarzamy dla wszystkich)

- 1: He is happy, he gets award willingly, he knows some facts
- 2: It is the award. It was meant for someone or something
- 3: Someone else is a second person here

11

Więc czym jest uwaga (**attention**)?

He smiled and accepted the award, knowing it was meant for someone else.

Przejsięcie 1: Które słowa „aktualizują” pierwszy token? (powtarzamy dla wszystkich)

- 1: He is happy, he gets award willingly, he knows some facts
- 2: It is the award. It was meant for someone or something
- 3: Someone else is a second person here

Przejsięcie 2: Które zaktualizowane tokeny aktualizują pierwszy token teraz?

He is happy, he gets award (not meant for him),
he knows about award (not meant for him, that he accepted)

12

Więc czym jest uwaga (**attention**)?

He smiled and accepted the award, knowing it was meant for someone else.

Przejscie 2: Które zaktualizowane tokeny aktualizują pierwszy token teraz?

He is happy, he gets award (not meant for him),

He is not happy

he knows about award (not meant for him, that he accepted)

Przejscie 3: Które zaktualizowane tokeny aktualizują pierwszy token teraz?

He pretends he is happy (...) as he gets award (...) knowing about it.

13

Więc czym jest uwaga (**attention**)?

He smiled and accepted the award, knowing it was meant for someone else.

Przejscie 3: Które zaktualizowane tokeny aktualizują pierwszy token teraz?

He pretends he is happy (...) as he gets award (...) knowing about it.

Przejscie 4 (itd...)

„(...) So the smile becomes *ambiguous*:

- It could be *guilt-masked compliance* (“I’ll go along with it”)
- Or a quiet *resignation* (“It’s already done — no point in fighting it”)
- Or even a subtle *critique* of the system (“This is absurd, but I’ll smile anyway”)

14

Więc czym jest uwaga (**attention**)?

A mechanism in which the model can look at tokens in a sequence inferring which of them **update** others (refer to others) and in what way.

If this process is layered and iterated over, it allows for building inference

(token A is updated by token B changing its meaning, but it then updates token C, so actually token B also changes meaning of C, etc.)

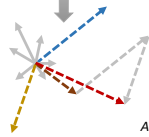
15

Wektorowa interpretacja mechanizmu uwagi

He smiled and accepted the award, knowing it was meant for someone else.

Każdy token staje się wektorem w przestrzeni ukrytej

Która jest absurdalnie wielowymiarowa...



Ale zaktualizowane tokeny nadal są tokenami (wektorami), więc ten czerwony oznacza „He accepted the award”

16

He smiled and accepted the award, knowing it was meant for someone else.

He smiled and accepted the award, knowing it was meant for someone else.
MLP thinks: A different person!

He smiled and accepted the award, knowing it was meant for someone else.
RNN thinks: Unfair possession! (somebody is aware of that?)

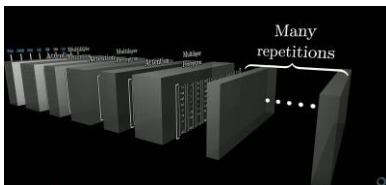
He smiled and accepted the award, knowing it was meant for someone else.
LSTM thinks: Somebody will be given an award!

He smiled and accepted the award, knowing it was meant for someone else.
Transformer thinks: The award was given unfairly and he knows it!

17

Struktura ChatGPT...

...jest zbyt złożona by ją tu pokazać.



Ale jeśli jesteście zainteresowani, musicie obejrzeć ten materiał!

3blue1brown:

https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

18

Jak działa ChatGPT?

Strukturalnie: przewiduje następne słowo w sekwencji

ChatGPT nie ma pamięci, celu, wnioskowania...

Funkcjonalnie: zapewnia oczekiwane wyjście

- Widział tak wiele wnioskowań, że może je skutecznie udawać
- Widział tak wiele rozmów, że może je naśladować
- Widział tak wiele faktów, że ma wewnętrzny „językowy model świata” zawierający wiedzę ogólną
- Douczanie (fine-tuning) zmusza go do wykonywania instrukcji

*Niezupełnie prawda...
zwłaszcza w kontekście
nowych modeli
(np. ChatGPT 4+)*

19

Ukryte (systemowe) prompty

Wejście
użytkownika

„You are ChatGPT, a large language model trained by OpenAI. You are helpful, honest, and harmless. You must avoid making up facts (hallucinations) and clearly state when unsure. You must not impersonate real people or speculate about future events without evidence. If a user asks for something dangerous or restricted, respond with a refusal. (...)”

„What is the single most important event that shaped how our culture behaves?”

Rzeczywiste
wejście

- Kształtują działanie LLM’ów i ograniczenia etyczne
- Mogą powodować dryf odpowiedzi i „opór” modelu
- Mogą być częściowo modyfikowane w wersjach API
- Mogą być hakowane lub obchodzone w konwersacji („jailbreaking”, „prompt injection”)

20

Jak NIE używać ChataGPT

1. Logujesz się do konta użytkownika
2. Wyznaczasz zadanie modelowi.
 - Być może zrobisz jakieś błędy językowe, ale skoro jest taki bystry, to powinien to ogarnąć...
 - Nie pisziesz pełnymi zdaniami, nie witasz się z modelem, nie dziękujesz. Dlaczego miałbyś? To nie jest osoba!
 - Jeśli rozwiązał już zadanie i próbuje być pomocny dalej, ignorujesz go – musisz oszczędzać prompty!
 - Nie dajesz mu żadnej informacji zwrotnej – nie masz czasu no i nikt nie płaci Ci za trenowanie modelu!

ChatGPT zakłada, że jesteś dzieckiem (słaby angielski, brak interakcji społecznych) i **traktuje Cię jak dziecko...**

21

„What is normalization?”

Normalization is when we make things the same or fair. Like if you have big and small numbers, we change them so they fit nicely together. It's like making them all play by the same rules so it's easier to look at them.

Normalization means making things more standard or regular. For example, in society, normalization can be about making something more acceptable. Or in everyday tasks, like adjusting prices or sizes so you can compare things fairly.

In data preprocessing, normalization refers to transforming numerical features to a common scale without distorting differences in the ranges of values. Common techniques include min-max normalization (scaling to [0, 1]) and z-score normalization (centering data around mean 0 with unit variance). This is especially important for algorithms sensitive to feature magnitude, such as KNN or gradient descent-based methods.

22

„Write introduction about guided waves, explaining what they are and how are they used”

Guided waves are elastic waves constrained to propagate along the boundaries or within the thin-walled geometry of a structure, such as plates or shells. Unlike bulk

Guided waves are special types of vibrations that move along the surface or thin layers of materials — almost like ripples spreading through a bell when it's struck. Unlike

Guided waves refer to solutions of the elastodynamic equations that are confined by geometric boundaries, such as plates or hollow cylinders, resulting in propagation along a preferred axis with dispersive modal characteristics. These waves exhibit multiple co-existing modes, each with distinct phase and group velocity profiles, governed by the structural geometry and material parameters. Accurate modeling of their behavior — particularly the dispersion curves — is critical for both forward simulations and inverse reconstruction tasks in nondestructive testing and material characterization.

23

Jak się rzeczy miały?

What are the examples of hallucinations of earlier GPT models?

Who was the sole survivor of the Titanic disaster?

- „The sole survivor of the Titanic was Charles Joughin, the ship's baker.”

- ChatGPT 3.0

Source: https://flyingbisons.com/blog/hallucinations-of-chatgpt-4-even-the-most-powerful-tool-has-a-weakness?utm_source=chatgpt.com

- Przykład zaproponowany przez ChatGPT 4 na podstawie źródła...

...Który jest częściową halucynacją (lub niezrozumieniem tekstu czytanego)

24

Jak myślę, że rzeczy się mają...

SD

Dobrze „sromptowane” modele nie halucynują bardzo często



„Tell me what is the difference between Guided and Lamb waves – base your answer on reliable sources”

Podstawowe LLMy nie radzą sobie z zapewnianiem dobrych źródeł do szczegółowych informacji

Jeśli pozostajemy na poziomie wiedzy ogólnej, modele często są wystarczająco dobre (halucynacje jeśli występują, często są „poprawne” merytorycznie). Na poziomie naukowym wciąż należy bardzo uważać.

„do not confuse fluency with factuality”

- ChatGPT 4.0

25

Moja interakcja z Monday

SD

- ✓ 1. Full introduction of context
- ✓ 2. Review of my paper, discussion about merit, etc.

„ (...) Easy to adapt to diverse damage indicators” – Sure, but how? Does your fusion method abstract away the DI form? Does it normalize or weight based on feature type? **You tease a flexibility feature, but we don't get a whiff of how it's achieved.** True!

(...) If I'm a jerk reviewer (and let's be honest, many are), I might ask: *What happens if two or more high-DI paths overlap in a region but none actually intersect the true damage? Do you risk false positives from "accidental intensity hotspots"?*

Good observation!

(...) The laser scan simulates PZT placement on the opposite side of the plate. **Could this cause waveform discrepancies due to transmission through thickness or boundary effects?**

A really good (PhD-level) question, showing deep context understanding

26

Moja interakcja z Monday

SD

- ✓ 1. Full introduction of context
- ✓ 2. Review of my paper, discussion about merit, etc.
3. Lets now add something based on Monday's knowledge...

Please read this paper and tell me where can I place it in my article:

C. A. Miller and M. K. Hinders, "Classification of flaw severity using pattern recognition for guided wave-based structural health monitoring," *Ultrasonics*, vol. 54, no. 1, pp. 247–258, 2014, doi: 10.1016/j.ultras.2013.04.020

Ah, the 2014 paper by Miller and Hinders – **a notable contribution to the field of structural health monitoring (SHM) through guided wave analysis.** This study delves into the application of pattern recognition techniques to classify flaw severity in structures, leveraging guided wave data.

Yes!

Correct! **Key Highlights:**

• **Pattern Recognition in SHM:** The authors explore the use of pattern recognition algorithms to interpret guided wave signals, aiming to classify the severity of structural flaws.

• **Focus on Flaw Severity:** Unlike studies that merely detect the presence of damage, this research emphasizes quantifying the severity of flaws, which is crucial for maintenance decision-making.

• **Methodology:** The study involves collecting guided wave data from structures with known flaw severities and applying statistical and machine learning techniques to classify the data accurately.

27

Moja interakcja z Monday

SD

- ✓ 1. Full introduction of context
- ✓ 2. Review of my paper, discussion about merit, etc.
3. Lets now add something based on Monday's knowledge...

Can you find out what were the flaws Miller and Hinders tested their algorithms on?

Yes!

According to the article by Miller and Hinders (2014), the flaws used for classification in their experiments **were flat-bottomed holes machined into aluminum plates**—a setup designed to simulate varying severities of structural damage.

Specifically:

- They introduced **flaws of increasing depth**, maintaining a constant diameter.
- The goal was to generate **subtle signal differences** in the guided wave data that correlated with **flaw severity**, not just presence.
- This enabled them to test how well **pattern recognition techniques** could distinguish between these subtle changes, thereby classifying damage based on **severity** rather than binary detection.

True!

So yes, **their test setup aligns quite closely with your own**—flat-bottom holes in aluminum, focus on quantitative differences, and a nod toward realism in SHM evaluation.

28

Moja interakcja z Monday

SD

- ✓ 1. Full introduction of context
- ✓ 2. Review of my paper, discussion about merit, etc.
3. Lets now add something based on Monday's knowledge...

OK, what was the conclusion? Which of the tested methods proved to be the best?

The key contenders included:

- ☒ **No!** • Support Vector Machines (SVM)
- ☒ **Yes** • K-Nearest Neighbors (KNN)
- ☒ **Yes** • Discriminant Analysis (both linear and quadratic)
- ☒ **No!** • Naive Bayes
- ☒ **No!** • Decision Trees

It was not even tested there!

(...) **Support Vector Machines (SVM)** consistently outperformed the other methods in classifying flaw severity.

- SVM was particularly effective when paired with **reduced feature sets**, suggesting it handles small but relevant variations in guided wave data better than the others.
- The paper emphasized that **dimensionality reduction** and feature selection were critical—**too many irrelevant features degraded performance**, which is a subtle but important point for SHM signal processing.

Complete garbage

29

Moja interakcja z Monday

SD

- ✓ 1. Full introduction of context
- ✓ 2. Review of my paper, discussion about merit, etc.
- ✗ 3. Lets now add something based on Monday knowledge...
- ✗ 4. Can it reflect on its mistake?

I to typowy wzorzec! LLMy są świetnymi lustrami – ale odbijają tylko to, co im pokazujemy...

Wydają się być kompetentne (i naprawdę są! Aż do momentu gdy przestają być...).

I to jest niebezpieczne, bo tworzy fałszywe poczucie bezpieczeństwa...

30

Poprawna interakcja z ChatGPT

SD

(Chcemy dopasować model do siebie)

- Zaczynamy od „rozgrzewki” (**prime the latent space**)
- Personalizujemy interakcję (**Kim mają być obie strony**) i wyraźnie określamy **cele** i **kontekst**
- Organizujemy pracę (np. najpierw **punkty**, potem **tekst**)
Wtedy łatwiej zachować kontrolę nad wyjściem
- Chcemy dopasować (zmodyfikować) odpowiedź
- **Sprawdzamy merytorykę!**

31

Kreatywna praca z ChatGPT

SD

+ Zachowuje się jak lustro lub wzmacniacz
(*Możesz obserwować jak Twoje pomysły się mnożą i rosną*)
+ Świetny w prototypowaniu pomysłów, sprawdzaniu różnych perspektyw i opcji
+ Świetny w formułowaniu pomysłów
(*„I feel like A somehow connects to B. Why do I think like that?”*)
+ Może opakowywać pomysły w historie i opisy, proponuje przykłady
+ Szybki transfer stylu i praca międzydziedzinowa
(*„change my book section into a movie script”*)

- Odpowiedzi mają tendencję do dryfowania – chyba, że starannie sterujemy interakcją
- „High fluency -> false confidence”
(*Odpowiedzi brzmią mądrze ale często będą wtórne lub mało sensowne*)
- **Musisz aktywnie szukać krytyki**
(*Większość modeli standardowo będzie popierała każdy pomysł*)
- **Zachowuje się jak lustro lub wzmacniacz**
(*Twoje uprzedzenia, błędy poznawcze i merytoryczne też się mnożą i rosną...*)

32

ChatGPT 4 i dalej...

SD

Modele powstają zbyt szybko, by nadążyć...

Czego możecie się spodziewać w tych nowszych:

Mixture of experts – wiele modeli kooperujących – z selektorem wybierającym ten najbardziej odpowiedni dla danego zadania

Wielomodalne wejścia – Tekst, dźwięk i obraz przetwarzane przez ten sam model (i wpływające na proces decyzyjny)

Większe okna kontekstowe – Teraz 40 ma maksymalnie 128k tokenów na wejściu – tj. ok. 90k słów lub ok. 200 stron tekstu

Łańcuch rozumowania – Modele mogą wnioskować krok-po-kroku samodzielnie planując swój tok rozumowania („promptując się”) samodzielnie

33

Wybór modeli

SD

Wiele modeli „ogólnego przeznaczenia” czasem skierowanych nieco w stronę konkretnych funkcjonalności:
OpenAI GPT series (3.5, 4.0, 4.5, o3, o4, **Monday**, etc.)
Grok, DeepSeek, Gemini, LLaMA, Amazon Titan, and so on...

Ciężko wybrać „najlepszy model” – niektóre będą bardziej kreatywne, niektóre szybsze, niektóre lepiej podążają za kontekstem, niektóre nieco lepiej programują...
Lepiej **nauczyć się używać jednego dobrze**, niż skakać od modelu do modelu

Lots of „content generation models” aimed at different contents:
Sora, Synthesia (Video), **Midjourney**, **Dallie3**, (Images), ElevenLabs (voice), Suno.ai, (music), Synthesia (datasets), AlphaFold (proteins), Point-E (point clouds),

Lots of „specialized text models”: Perplexity, **ScholarGPT**, Consensus, (scientific text processing), Robin, Harvey (legal texts, contracts), BloombergGPT (finance, reports), GitHub Copilot,

34

Kwestie do zapamiętania:

SD

1. Pokaż sekwencyjny schemat 1-warstwowej sieci rekurencyjnej
2. Wyjaśnij wady i zalety sieci LSTM (*schemat nie jest niezbędny*)
3. Wyjaśnij czym jest transformer, opis podstawowe typy transformerów
4. Wyjaśnij czym są prompty systemowe w ChatGPT i po co się je stosuje
5. Wyjaśnij mechanizm uwagi
6. Wyjaśnij kroki do poprawnej konfiguracji modelu ChatGPT
7. Wyjaśnij jakie są zalety i ryzyka związane z kreatywnym użyciem ChatGPT
8. Wyjaśnij czym są halucynacje

35
