

Rok 2017: Artykuł, który zredefiniował fundamenty NLP

Attention Is All You Need

Manifest

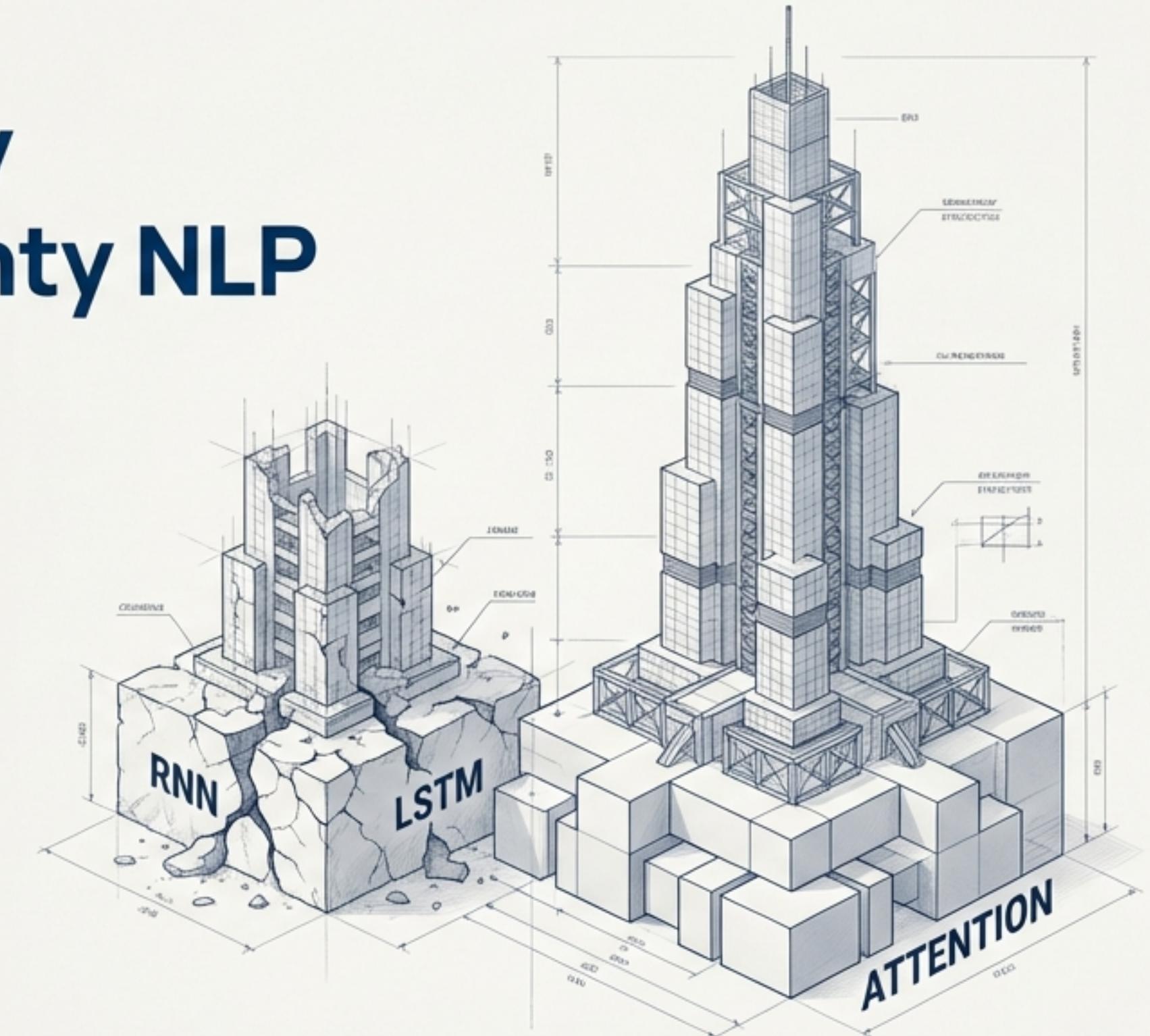
Tytuł pracy to nie tylko opis, ale deklaracja ideologiczna. Autorzy proponują radykalne uproszczenie, odrzucając dekady rozwoju modeli rekurencyjnych i konwolucyjnych.

Fundamentalne Pytanie

Co, jeśli model językowy nie musi przetwarzać zdania słowo po słowie? Co, jeśli może spojrzeć na nie w całości, na raz?

Nowa Architektura

Transformer. Nazwa, która stała się synonimem nowoczesnego AI. Jego siła nie leży w rekurencji czy konwolucjach, ale wyłącznie w mechanizmie uwagi.



“We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.”

Problem: Modele sekwencyjne osiągnęły swój limit

Główne Ograniczenia RNN/LSTM

- **Przetwarzanie Krok po Kroku:** Aby zrozumieć 20. słowo, model musi przetworzyć 19 poprzednich. Ta "z natury sekwencyjna natura" uniemożliwia efektywną paralelizację obliczeń.
- **Problem Długiego Zasięgu:** Informacje z początku zdania ulegają zniekształceniu w miarę przetwarzania. Modele miały trudność z nauką zależności między odległymi słowami.
- **Ograniczona Ścieżka Informacji:** Długość ścieżki, jaką sygnał musi przebyć między dwoma pozycjami, rośnie liniowo ($O(n)$) wraz z długością sekwencji.

Nawet ulepszenia takie jak LSTM, GRU czy ConvS2S były jedynie “plastrami” na fundamentalny problem.
Potrzebne było nowe podejście.



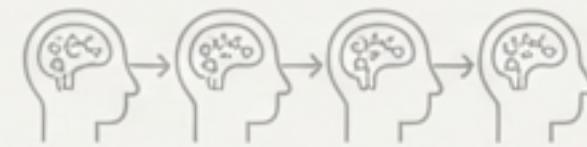
Przetwarzanie Sekwencyjne

Porównanie Złożoności

Typ Warstwy	Złożoność na Warstwę	Operacje Sekwencyjne	Maksymalna Długość Ścieżki
Recurrent	$O(n * d^2)$	$O(n)$	$O(n)$

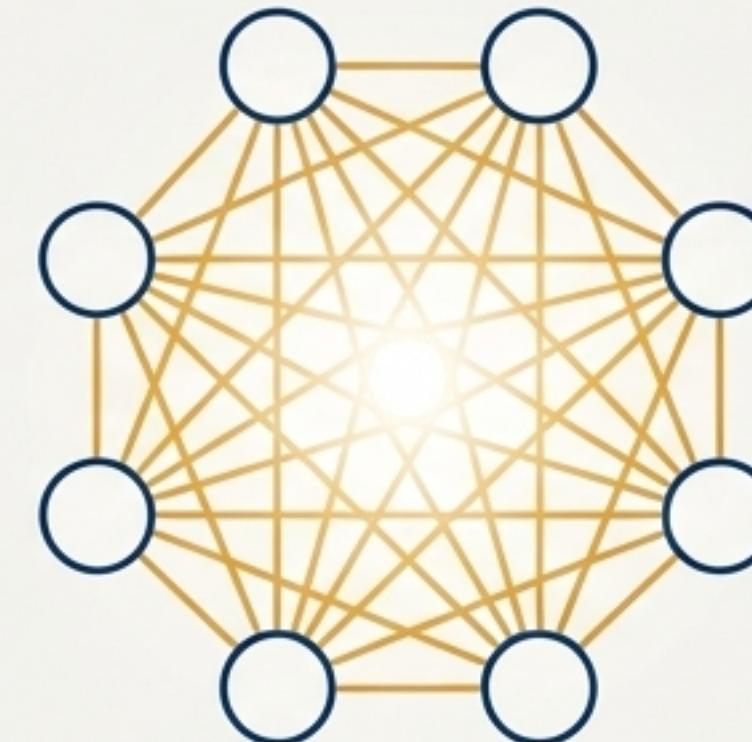
Przełom: Self-Attention pozwala każdemu słowu widzieć wszystkie inne jednocześnie

Zamiast przetwarzać słowa po kolei, mechanizm self-attention oblicza reprezentację całej sekwencji jednocześnie.



RNN: $O(n)$ ścieżka

Zmiana paradygmatu



Self-Attention: $O(1)$ ścieżka

Każde słowo może bezpośrednio "zapytać" każde inne: "Jak bardzo jesteś dla mnie istotne w tym kontekście?".

Nie ma pośredników.

****Kluczowa Korzyść**:** Długość ścieżki między dowolnymi dwoma pozycjami w sieci jest stała. Problem zależności dalekiego zasięgu zniką. To umożliwia ogromną paralelizację obliczeń.

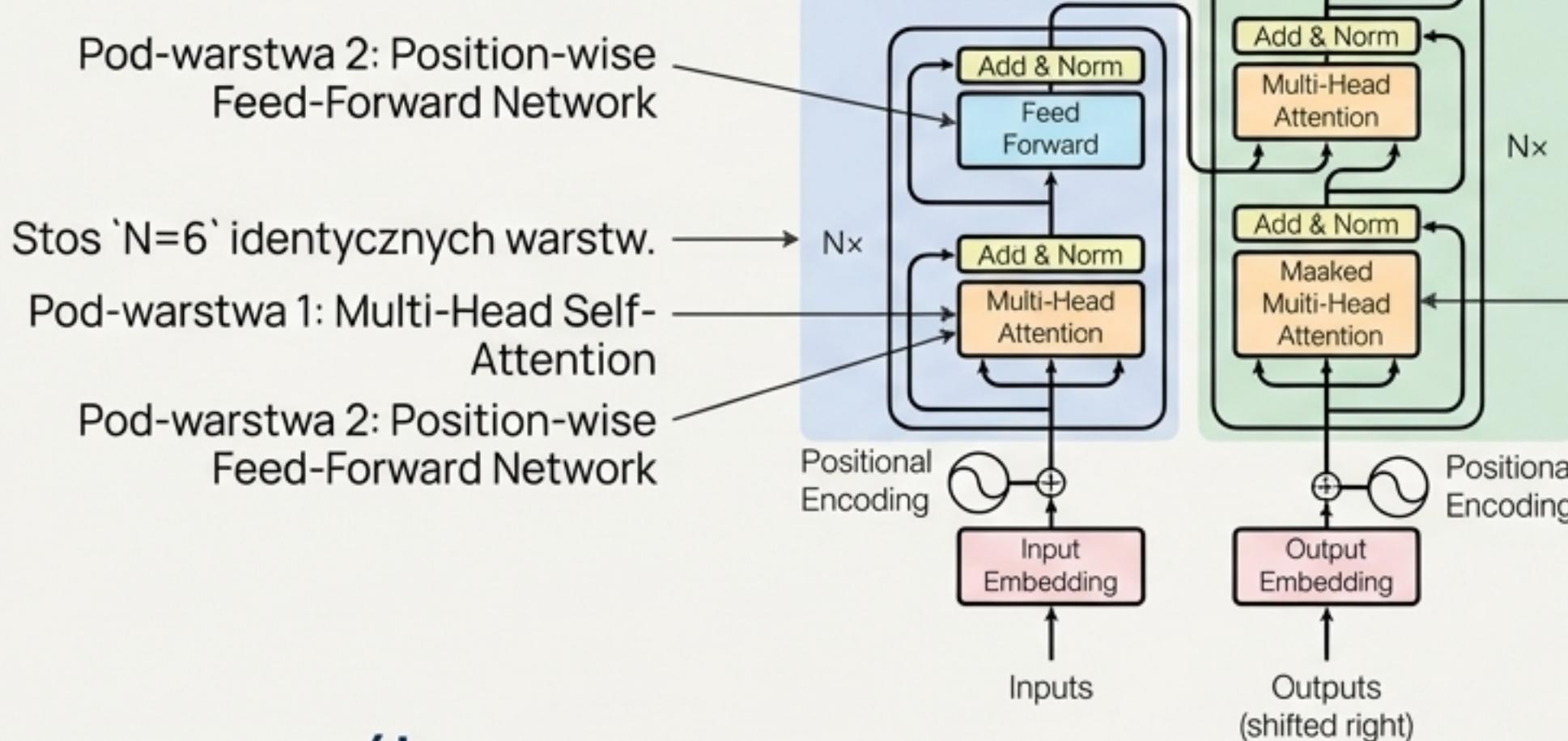
Porównanie Złożoności

Typ Warstwy	Złożoność na Warstwę	Operacje Sekwencyjne	Maksymalna Długość Ścieżki
Recurrent	$O(n * d^2)$	$O(n)$	$O(n)$
Self-Attention	$O(n^2 * d)$	$O(1)$	$O(1)$

Anatomia rewolucji: Architektura modelu Transformer

Enkoder

- Zadanie: Przekształca sekwencję wejściową w ciągłą, bogatą w informacje reprezentację numeryczną.



Dekoder

- Zadanie: Generuje sekwencję wyjściową słowo po słowie, bazując na reprezentacji z enkodera.

Również stos 'N=6' warstw.

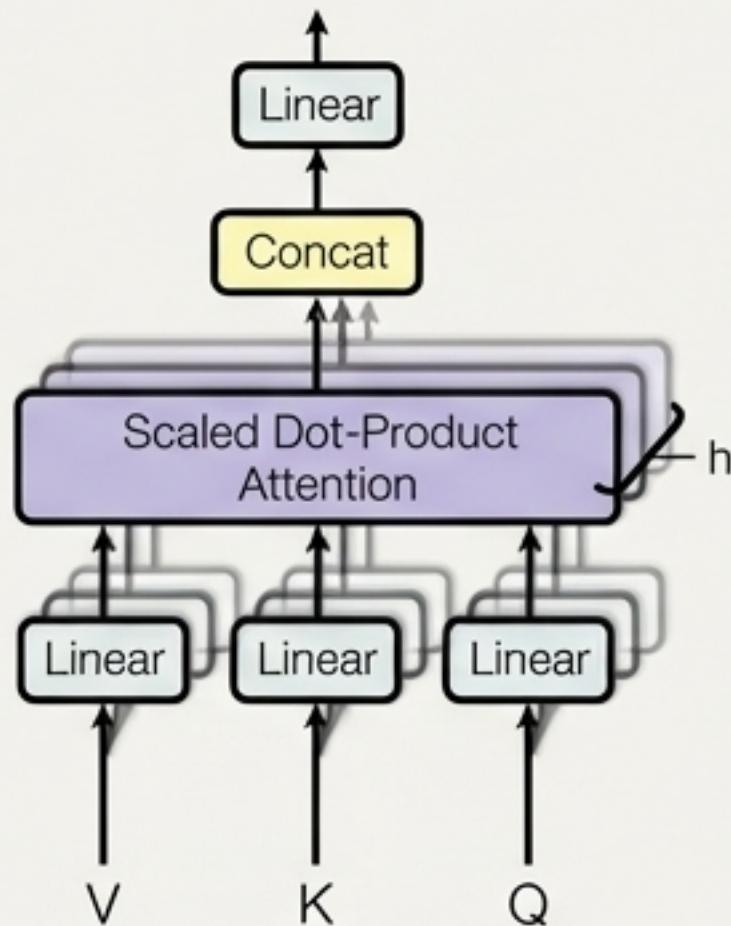
Dodatkowa pod-warstwa 3: Masked Multi-Head Attention, która pozwala zwracać uwagę na wyjście enkodera.

Kluczowe szczegóły

Wokół każdej pod-warstwy zastosowano połączenia rezydualne ('Add') i normalizację warstwową ('Norm'), co jest kluczowe dla treningu głębokich modeli.

Zespół specjalistów: Jak działa Multi-Head Attention

Mechanizm (h=8)



1. Oryginalne wektory Query, Key i Value są liniowo transformowane ' h ' razy do niższych wymiarów.
2. Mechanizm uwagi (Scaled Dot-Product Attention) jest stosowany równolegle dla każdej z ' h ' wersji.
3. Wyniki z ' h ' głów są konkatenowane i ponownie transformowane, tworząc ostateczne wyjście.

`'d_model' = 512, ' h ' = 8, więc wymiar każdej głowy to ' d_k ' = ' d_v ' = 64.`

Korzyść: Różne Perspektywy



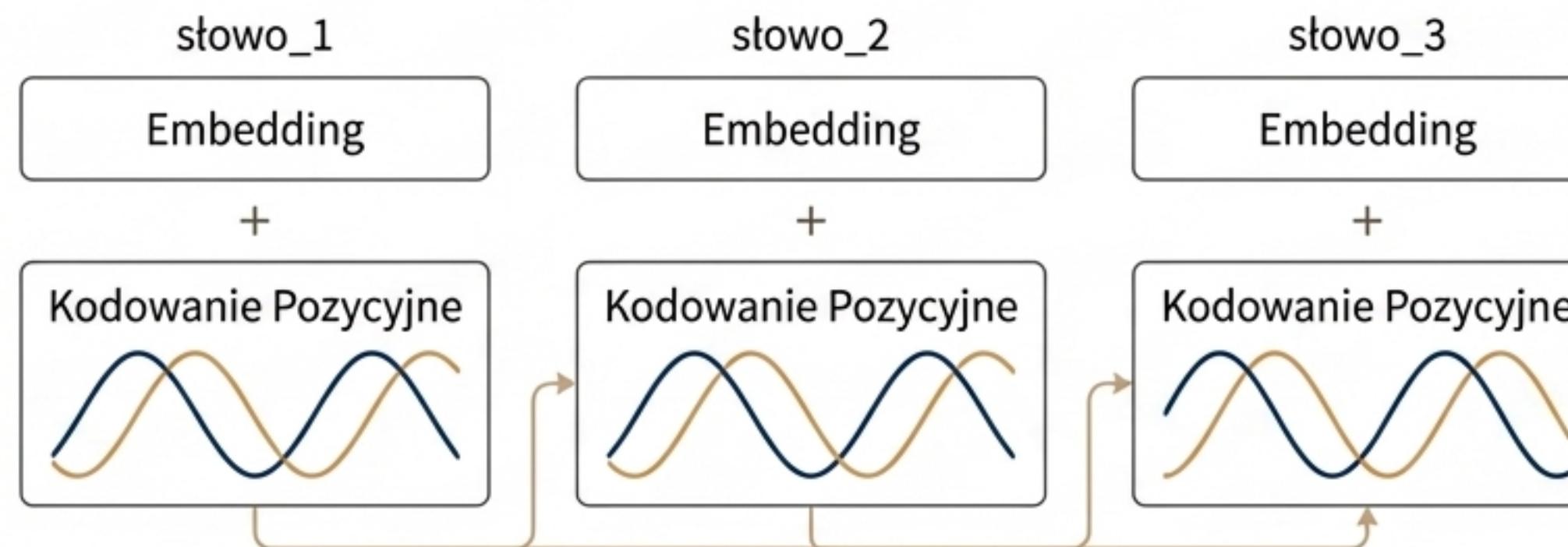
“ “Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions.” ”

Jak zachować porządek bez sekwencji: Magia kodowania pozycyjnego

Problem

W architekturze opartej wyłącznie na self-attention, zdania "pies goni kota" i "kot goni psa" wyglądałyby tak samo, ponieważ model nie ma wbudowanego pojęcia kolejności.

Rozwiązanie: Dodanie "wektora-adresu" do każdego słowa



$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

Zalety tego podejścia

- Każda pozycja ma unikalne kodowanie.
- Model może łatwo uczyć się zależności od względnej pozycji.
- Pozwala modelowi generalizować na sekwencje dłuższe niż te widziane podczas treningu.

Werdykt: Transformer deklasuje konkurencję w tłumaczeniu maszynowym

Benchmark: WMT 2014 Machine Translation (BLEU score)

Angielski -> Niemiecki (EN-DE)



Nowy Rekord Świata

Angielski -> Francuski (EN-FR)

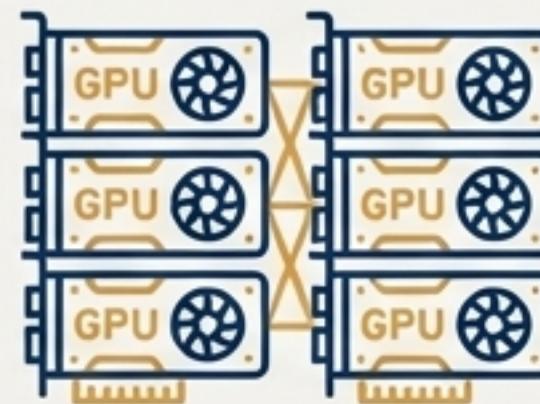
41.8 BLEU

Najwyższy wynik dla pojedynczego modelu, deklasujący poprzednie systemy.

To nie była kosmetyczna poprawa. To był nokaut.

Druga rewolucja: radykalne skrócenie czasu treningu

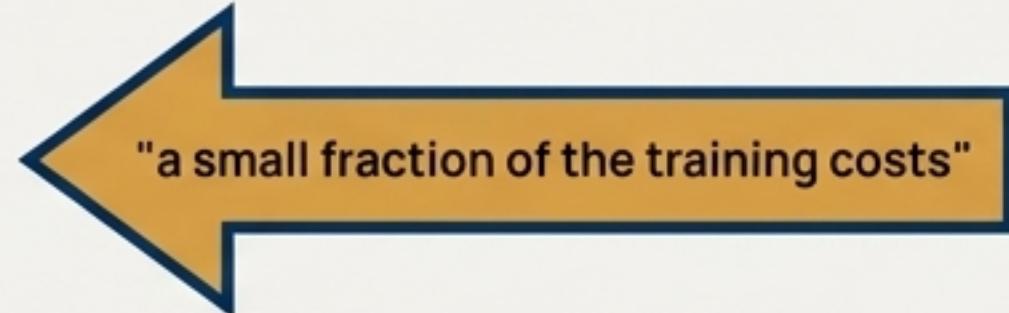
Transformer (big)



Poprzednie modele SOTA



3.5 Dnia



Wiele tygodni

Implikacje

Demokratyzacja Badań

Po raz pierwszy, osiągnięcie wyników SOTA nie wymagało dostępu do gigantycznych farm serwerów przez długi czas.

Przyspieszenie Innowacji

Krótszy cykl treningu pozwolił na znacznie szybsze eksperymentowanie, otwierając drzwi do eksploracji modeli na skalę, która wcześniej była nieosiągalna.

Dowód uniwersalności: Zastosowanie w parsingu składniowym

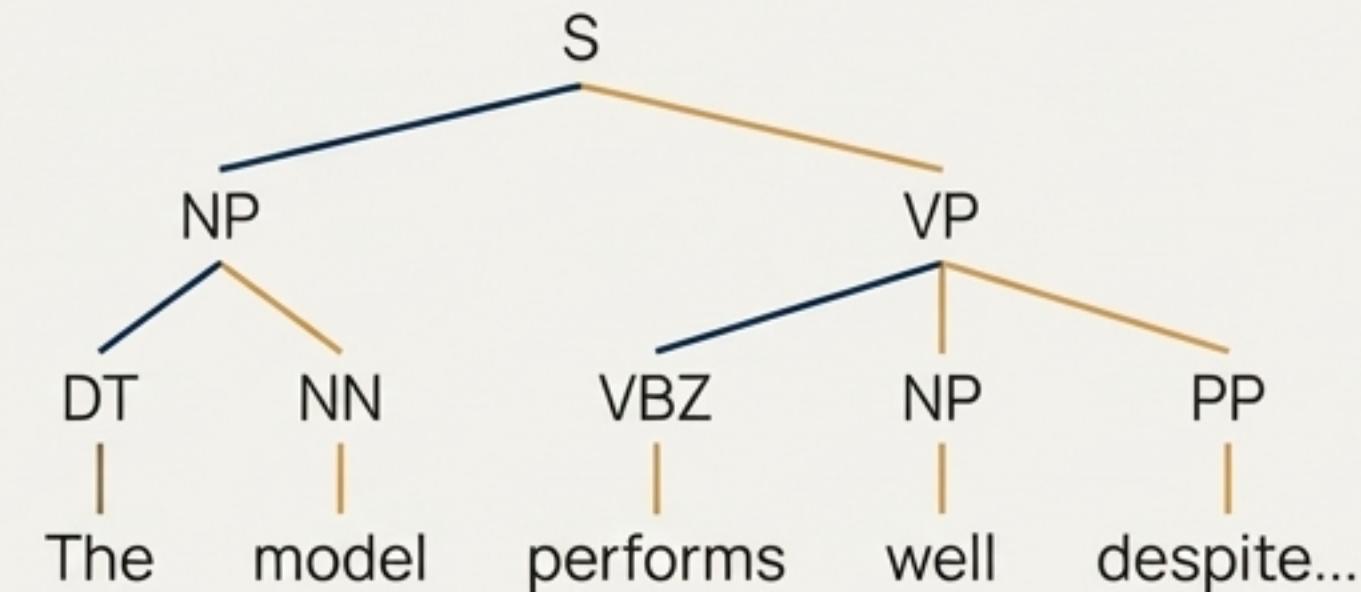
Nowe zadanie: English constituency parsing

Autorzy przetestowali model na zadaniu analizy struktury gramatycznej zdania, aby sprawdzić jego zdolność do generalizacji.

Wyzwania:

- Wyjście ma silne ograniczenia strukturalne
 - Wyjście jest często znacznie dłuższe niż wejście.
 - Modele RNN nie radziły sobie dobrze w tym zadaniu.

Zaskakujące wyniki (F1 Score)



Tylko na 40 tys. zdań

Transformer: 91.3

BerkeleyParser: 90.4

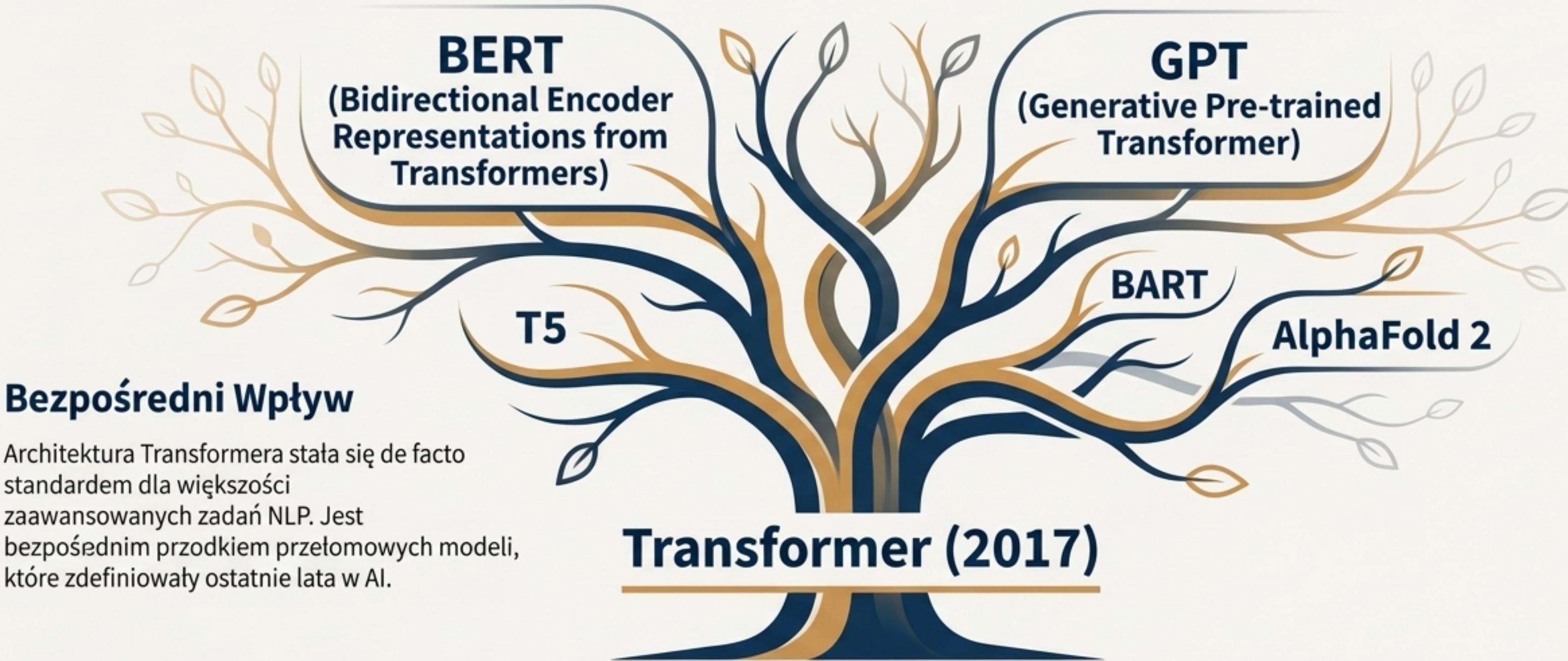
Z dodatkowymi danymi

Transformer: 92.7

Poprzednie modele SOTA: <92.2

Cytat z artykułu: "...despite the lack of task-specific tuning our model performs surprisingly well..."

Dziedzictwo 'Attention Is All You Need': Fundament nowoczesnej sztucznej inteligencji



Manifest z 2017 roku okazał się proroczy. Uwaga stała się kluczowym mechanizmem napędzającym najbardziej zaawansowane systemy AI, z których korzystamy dzisiaj. To jedna z najważniejszych prac naukowych w historii informatyki.