



POLITECHNIKA WROCŁAWSKA
WYDZIAŁ PODSTAWOWYCH PROBLEMÓW TECHNIKI

Praca inżynierska

**Rozwój mobilnej aplikacji do rozpoznawania
języka migowego z wykorzystaniem
transformerów**

Development of mobile application for sign language recognition
using transformers

Autor: Maciej Grzesik

Opiekun: Prof. uczelni dr hab. Sebastian Kraszewski

Wrocław 2024

Spis treści

1	Wprowadzenie	2
1.1	Słowa kluczowe	2
1.2	Cel pracy	2
1.3	Związek z inżynierią biomedyczną	2
2	Teoria	3
2.1	Zasada działania	3
2.2	Architektura	5
	Bibliografia	7

Rozdział 1

Wprowadzenie

1.1 Słowa kluczowe

Rozpoznawanie języka migowego, aplikacja mobilna, flutter, uczenie głębokie, transformer, wizja komputerowa

1.2 Cel pracy

Celem pracy jest stworzenie krosplatformowej aplikacji mobilnej w środowisku Flutter opartym o język programowania Dart, która wykorzystuje architekturę transformerów do rozpoznawania i tłumaczenia języka migowego na tekst w czasie rzeczywistym.

1.3 Związek z inżynierią biomedyczną

Osoby z upośledzeniem słuchu zmagają się z wykluczeniem społecznym, ze względu na brak możliwości komunikacji za pomocą konwencjonalnej mowy. Na podstawie badań Pauliny Malczewskiej wynika, że aż 90% ankietowanych osób głuchych doświadcza alienacji oraz dyskryminacji ze strony osób słyszących [1].

Bariery komunikacyjne między osobami głuchoniemymi a tymi, którzy nie znają języka migowego, stanowią znaczący problem społeczny i medyczny. Przyczynia się to do izolacji społecznej, trudności w dostępie do usług zdrowotnych i ogólnie niższej jakości życia.

Stworzenie aplikacji wspierającej osoby głuche nie tylko ułatwi komunikację, ale także przyczyni się do zwiększenia inkluzywności społecznej, zapewniając równe szanse i redukując poziom dyskryminacji.

Rozdział 2

Teoria

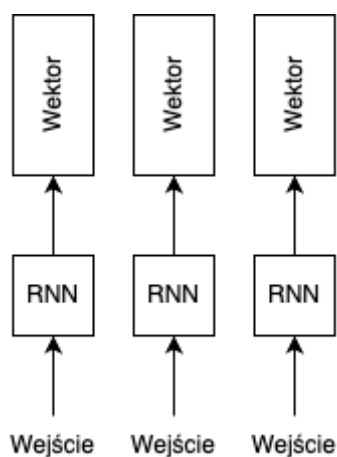
Transformery to zaawansowane modele uczenia maszynowego, szeroko stosowane w przetwarzaniu języka naturalnego, wizji komputerowej, ale także w audio i przetwarzaniu multimedialnym.

W odróżnieniu od poprzednio stosowanych modeli uczenia maszynowego opierających się na podejściu rekurencyjnym i mechanizmie uwagi, naukowcy pracujący dla firmy Google zaproponowali rozwiązanie wykorzystujące jedynie mechanizm uwagi. Skutkiem tego podejścia jest znaczne przyspieszenie czasu uczenia modelu przy zachowaniu wysokiej precyzji [2].

2.1 Zasada działania

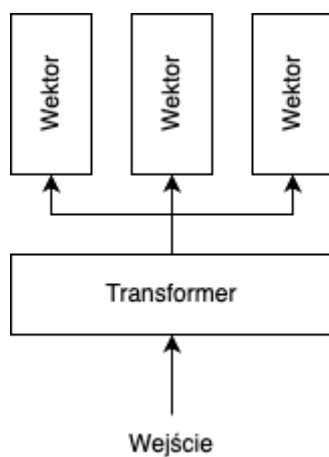
Działanie transformera opiera się na warstwie kodującej oraz na warstwie dekodującej. Zadaniem warstwy kodującej jest zakodowanie argumentów wejściowych (np. tekstu, obrazów) do postaci numerycznej. Warstwa dekodująca, wykorzystuje przekazane z warstwy kodującej zakodowane informacje do przetworzenia ich na wartości wyjściowe (np. do tekstu czy wideo).

Istotnym elementem, który różnicuje transformery od rekurencyjnych modeli uczenia maszynowego jest mechanizm uwagi. W tradycyjnych modelach rekurencyjnych sieci neuronowych (RNN) dane przetwarzane są sekwencyjnie tzn., że są analizowane krok po kroku (patrz 2.1) [3] co skutkuje ograniczeniami m.in. w równoległości przetwarzaniu danych.



Rysunek 2.1: Przepływ informacji wejściowych w rekurencyjnych sieciach neuronowych

Rozwiązania inżynierskie wykorzystane w transformerach pozwalają modelom na przetwarzanie wszystkich elementów sekwencji jednocześnie (patrz [2.2](#)).



Rysunek 2.2: Przepływ informacji wejściowych w transformerach

W praktyce oznacza to, że uczenie transformerów przebiega w określonej stałej ilości $O(1)$ sekwencyjnie wykonywanych operacji, modele RNN wymagają natomiast $O(n)$ sekwencyjnie wykonywanych operacji co sprawia, że są mniej wydajne przy dłuższych sekwencjach.

2.2 Architektura

Zanim dane przetwarzane przez transformer zostaną przekazane do kodera muszą zostać wstępnie przetworzone. Proces ten można podzielić na dwa kroki:

- Input Embedding polega na zamienieniu danych wejściowych na przystępne dla obliczeń komputerowych wektory.
- Positional Encoding składa się z wektorów, których zadaniem jest nadanie kontekstu na podstawie pozycji danych w sekwencji. Jest to istotne, gdyż z punktu widzenia maszyny nie jest w stanie odróżnić ona informacji użytych w różnym znaczeniu.

Sam koder składa się z dwóch elementów:

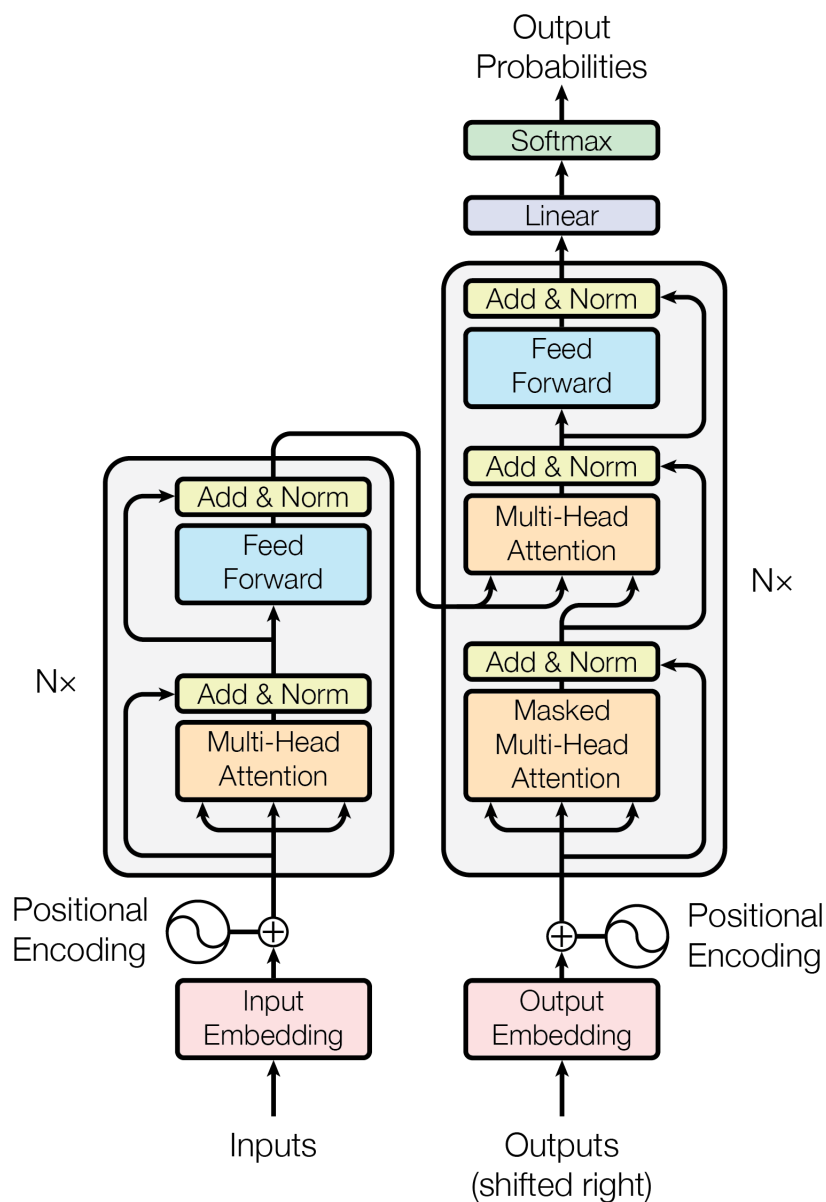
- Multi-Head Attention polega na obliczeniu wektora uwagi dla poszczególnych informacji przekazanych do tej warstwy. W tym etapie obliczany jest tzw. wektor uwagi za pomocą którego określa się wagę poszczególnych informacji. W przypadku przetwarzania wideo identyfikowane są kluczowe momenty takie jak m.in. gesty poprzez nadanie wag poszczególnym klatkom wideo w zależności od ich znaczenia. Ignorowane są mniej istotne informacje lub szum.
- Feed Forward, w tym etapie obliczone wektory przekazywane są do sieci MLP (Multi Layer Perceptron), która używana jest dla każdego wektora uwagi.

Tak przygotowane dane zostają przekazane do dekodera, którego zadaniem jest przewidywanie kolejnych słów czy obrazów. W tym procesie udział biorą poszczególne bloki:

- Embedding proces ten przebiega dokładnie tak samo jak w przypadku warstwy kodującej. Dane zamieniane są na wartości numeryczne w postaci macierzy.
- Positional Encoding również odbywa się w taki sam sposób jak w przypadku warstwy kodującej. Obliczane są wektory nadające kontekst informacjom.
- Masked Multi-Head Attention polega na zamaskowaniu, czyli przemnożeniu przez macierz zer.
- Multi-Head Attention with encoder łączy wektor wyjściowy z warstwy enkodującej z wektorem z poprzedniego kroku ze sobą. Podczas tego etapu sprawdzane jest w jakim stopniu każdy wektor jest ze sobą powiązany.

- Feed Forward jest to sieć, której zadaniem jest uproszczenie tłumaczenia wektora, aby łatwiej można było przerobić transformerowi wyniki parowań. Następnie w etapie linear layer przekształcane są wyniki mające na ten sam wymiar co dane wejściowe. Następnie przy użyciu funkcji softmax zmieniają się wyniki prawdopodobieństwa.

Poszczególne bloki architektury wraz z przepływem informacji w modelu zostały zaprezentowane na schemacie 2.3 [2].



Rysunek 2.3: Schemat blokowy architektury transformera

Bibliografia

- [1] P. Malczewska. Izolacja społeczna osób z uszkodzonym słuchem jako wspólny obszar badań pedagogiki i antropologii. *Pedagogika a etnologia i antropologia kulturowa. Wspólne obszary badań*, page 128, 2011. [2](#)
- [2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. [3](#), [6](#)
- [3] M. Mamczur. Czym jest i jak działa transformer sieć neuronowa?, March 2020. Dostęp: 23 maj 2024. [3](#)