



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 asda asds

Opiekun: Prof. uczelni dr hab. Sebastian Kraszewski

Wrocław 2024

Moim ...

Spis treści

1	Wprowadzenie	7
1.1	Motywacja	7
1.2	Cel	8
1.3	Zakres	8
2	Teoria	9
2.1	Zasada działania	9
2.2	Architektura	11
	Bibliografia	13

Rozdział 1

Wprowadzenie

1.1 Motywacja

Wykluczenie społeczne osób głuchoniemych stanowi poważny problem w społeczeństwie, wynikający z braku możliwości komunikacji werbalnej w sferze publicznej. Uszkodzenie słuchu, szczególnie w życiu codziennym, powoduje trudności w komunikacji z otoczeniem, które opiera się na konwencjonalnej mowie. Rodzi to problem z możliwością partycypacji w prostych aktywnościach poczynając od uczestniczenia w życiu społecznym do m.in. jakości otrzymywanych usług medycznych.

Należy zaznaczyć, iż znajomość języka migowego wśród ludzi zdrowych ogranicza się do sytuacji w których osoby im bliskie są dotknięte problemem uszkodzonego słuchu. Język ten, będący podstawową formą komunikacji dla osób głuchoniemych nie jest szeroko nauczany a jego znajomość często wymaga zaangażowania się w dodatkowe kursy tudzież szkolenia, nierzadko płatne.

Na podstawie badań Pauliny Malczewskiej wynika, że aż 90% ankietowanych głuchoniemych doświadcza alienacji oraz dyskryminacji ze strony słyszących [1]. Zjawiska te niezaprzeczalnie przyczyniają się do przewlekłego obniżenia nastroju, lęku przed byciem postrzeganym przez społeczeństwo oraz ogólnym spadkiem poczucia własnej wartości i samooceny. Te aspekty stawiają solidny grunt pod rozwój chorób psychicznych tj. depresja, epizody nastroju depresyjnego, zaburzenia adaptacyjne czy fobii społecznej co potwierdzają poszczególne wytyczne diagnostyczne zawarte w DSM-5 [2].

Istotnym jest zaadresowanie tego problemu za pomocą aplikacji wspierającej osoby głuchonieme. Nie tylko ułatwi to komunikację, ale także przyczyni się do zwiększenia inkluzyjności społecznej, zapewniając równe szanse i redukując poziom dyskryminacji.

1.2 Cel

Głównym celem pracy jest stworzenie algorytmu opartego na transoformerach czyli na architekturze głębokiego uczenia maszynowego, którego zadaniem będzie klasyfikacja znaków języka migowego. Istotnym jest zasięgnięcie do rozwiązań opartych na uczeniu maszynowym gdyż rozwiązania oparte na algorytmice nie sprawdzają się w przypadkach dotyczących widzenia komputerowego, które jest znaczącym elementem tej pracy. [elaborate on this] Umożliwi to tłumaczenie języka migowego na tekst w czasie rzeczywistym, co ułatwi komunikację osobom głuchoniemym.

Sam algorytm zaimplementowany zostanie w aplikacji mobilnej docelowo dedykowanej na smartfony z systemem operacyjnym Android oraz iOS. Możliwe jest to dzięki wykorzystaniu środowiska Flutter, które pozwala na tworzenie oprogramowania opartego o język Dart, a następnie na kompilowanie kodu przy wykorzystaniu natywnych narzędzi kompilacyjnych (Android Studio dla systemu Android; Xcode dla systemu iOS).

1.3 Zakres

W ramach pracy inżynierskiej zaprojektowany zostanie przyjazny interfejs użytkownika, który umożliwi proste korzystanie z aplikacji zarówno przez osoby głuche, jak i słyszące. Wykorzystanie środowiska Flutter pozwoli na implementację interaktywnego interfejsu użytkownika przy wykorzystaniu wbudowanych funkcjonalności. Dodatkowo zostanie wyeliminowana potrzeba pisania kodu w natywnym dla danego środowiska języku, co zapewnia aplikacji możliwość działania na różnych mobilnych systemach operacyjnych.

Zostanie również przeprowadzona implementacja oraz trening modelu. W tym etapie projektu zostaną zastosowane odpowiednie algorytmy, które umożliwią skuteczne uczenie modelu na podstawie zebranych danych dotyczących gestów języka migowego. Proces uczenia będzie obejmował zarówno fazę wstępną, w której model będzie dostosowywany do charakterystyki danych, jak i fazę walidacji, w której oceni się jego dokładność i zdolność do generalizacji gestów języka migowego. Istotnym jest również przeprowadzenie testów funkcjonalności modelu w warunkach rzeczywistych aby potwierdzić jego poprawne działanie lub wprowadzanie ewentualnych poprawek.

Rozdział 2

Teoria

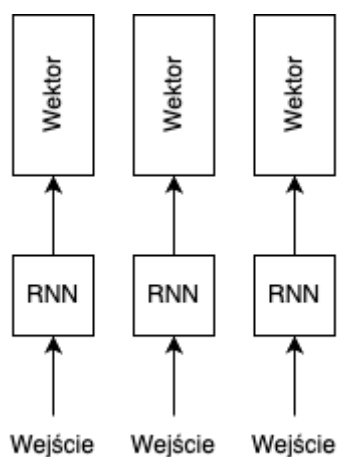
Transformery to zaawansowane modele uczenia maszynowego, szeroko stosowane w przetwarzaniu języka naturalnego, wizji komputerowej, ale także w audio i przetwarzaniu multimedialnym.

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 [3].

2.1 Zasada działania

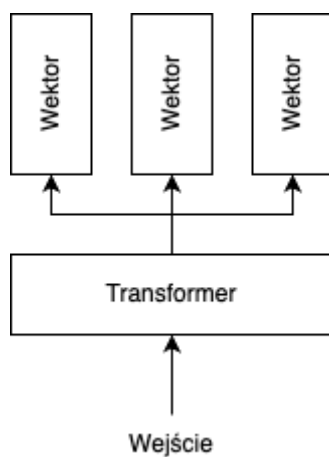
Działanie transformera opiera się na warstwie kodującej oraz na warstwie dekodującej. Zadaniem warstwy kodującej jest zakodowanie argumentów wejściowych (np. tekstu, obrazów) do postaci numerycznej. Warstwa dekodująca, wykorzystuje przekazane z warstwy kodującej zakodowane informacje do przetworzenia 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) [4] 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:

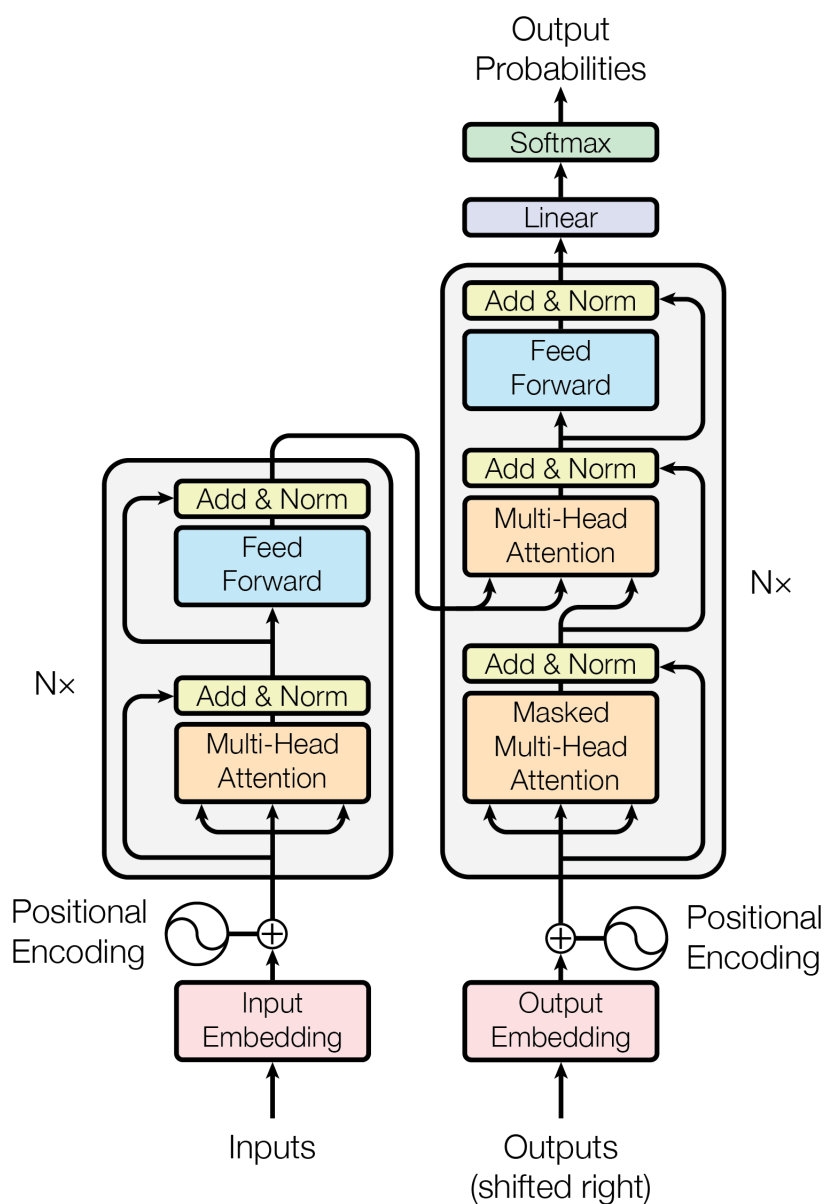
- Multi-Head Attention polega na obliczeniu wektora uwagi dla poszczególnych informacji przekazanych do tej warstwy. W tym etapie obliczany jest tzw. wektor uwagi za pomocą którego określa się wagę poszczególnych informacji. W przypadku przetwarzania wideo identyfikowane są kluczowe momenty takie jak 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 [3].



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. [7](#)
- [2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Publishing, 2013. [7](#)
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. [9](#), [12](#)
- [4] M. Mamczur. Czym jest i jak działa transformer sieć neuronowa?, March 2020. Dostęp: 23 maj 2024. [9](#)