Transforméry a iné výpočtové metódy pri spracovaní medicínskych signálov

Analýza a návrh riešenia

Daniela Pillárová

1 Úvod

Táto diplomová práca tematicky nadväzuje na moju predchádzajúcu bakalársku prácu, ktorá sa zaoberala Predikciou prognózy pacientov po zástave srdca pomocou analýzy EEG záznamov metódami strojového učenia. V rámci bakalárskej práce bolo vytvorených niekoľko základných modelov predikcie stavu pacienta založených primárne na Náhodných lesoch a Ensemble modeloch. Dosiahnuté výsledky, ako aj iné dostupné štúdie, poukázali na potenciál využitia umelej inteligencie pri hodnotení neurologického stavu komatózneho pacienta, no zároveň poukázali na potrebu ďalšej, hlbšej analýzy charakteristík EEG signálov.

V tejto nadväzujúcej diplomovej práci sa zameriavam na pokročilejšie spracovanie EEG signálov pacientov po zástave srdca. Hlavným cieľom je využitie modelov hlbokého učenia na analýzu týchto údajov, pričom sa plánujem venovať najmä architektúram, ktoré sú schopné efektívne pracovať s veľkými množstvami dát. Okrem klasických neurónových sietí budem osobitný dôraz klásť na modely typu transformér, ktoré v poslednom období dosahujú výborné výsledky aj pri spracovaní sekvenčných a časových údajov.

2 Charakteristika problému

Problém, ktorému sa táto práca venuje, súvisí s predikciou neurologického stavu pacientov, ktorí po zástave srdca upadli do kómy. V takýchto prípadoch je pacientovi často poskytnutá včasná resuscitácia, čo vedie k obnoveniu krvného obehu v priebehu niekoľkých minút od incidentu. Napriek tomu však dochádza k závažnému narušeniu činnosti centrálneho nervového systému a pacient zostáva v bezvedomí.

Po prevoze do nemocnice je pacient dlhodobo monitorovaný, najmä prostredníctvom EEG záznamov, ktoré poskytujú informácie o elektrickej aktivite mozgu. V tomto štádiu sú lekári často konfrontovaní s otázkami zo strany rodiny a blízkych ohľadom stavu a perspektív pacienta. Odpovedať na tieto otázky však býva náročné, keďže neurologická prognóza v akútnej fáze nie je vždy

jednoznačná.

Ako pomocný nástroj pri hodnotení stavu pacienta sa v klinickej praxi využíva tzv. Cerebral Performance Category (CPC) skóre, ktoré hovorí o možnom budúcom neurologickom stave, povedzme za 6 mesiacov. CPC tak rozdeľuje pacientov do piatich kategórií od 1 – dobrý neurologický stav až po 5 – mozgová smrť (tab. 1). Toto skóre sa určuje na základe celkového klinického obrazu vrátane EEG záznamov počas prvých hodín až dní po incidente.

CPC skóre	**Popis**
1	Dobrý mozgový stav – pacient je plne pri vedomí, schopný samostatného života, môže mať minimálne neurologické deficity.
2	Mierne poškodenie – pacient je pri vedomí a funkčne nezávislý, no s miernym neurologickým postihnutím (napr. poruchy pamäti, reči, koordinácie).
3	Ťažké neurologické postihnutie – pacient je pri vedomí, ale trvale závislý od iných osôb v každodennom živote.
4	Kóma alebo vegetatívny stav – pacient nevykazuje vedomie o okolí, nereaguje zmysluplne na podnety.
5	Mozgová smrť – úplná strata mozgovej činnosti, ireverzibilný stav.

Tabuľka 1: Cerebral Performance Category (CPC) skóre – klinická klasifikácia neurologického stavu.

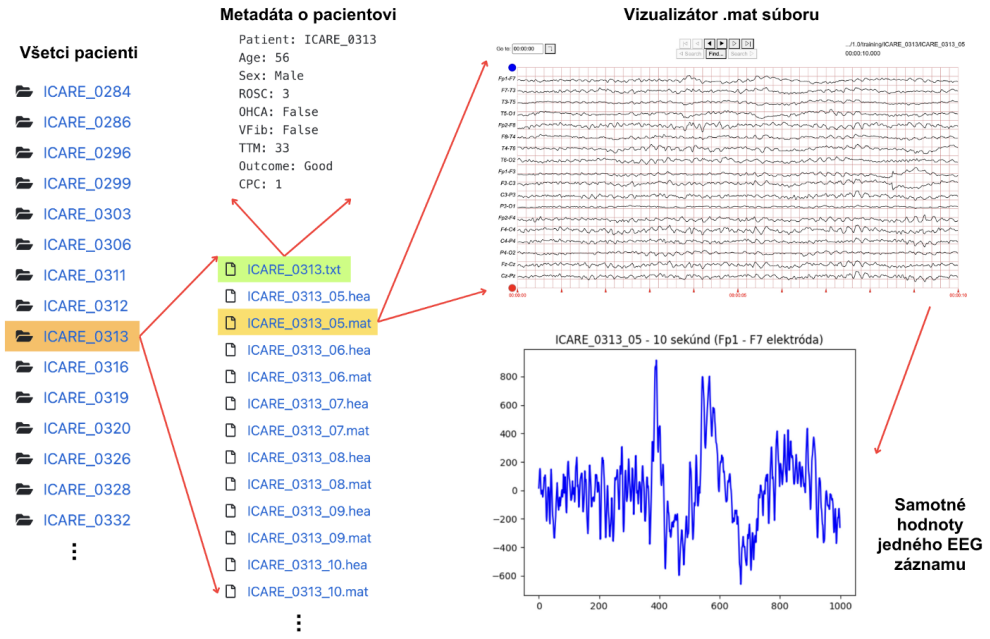
EEG záznamy pacientov po zástave srdca obsahujú viaceré charakteristiky, ktoré napovedajú k správnemu určeniu CPC skóre. Medzi najčastejšie sledované vzory patrí:

- redukovaná voltáž – zníženie amplitúdy elektrickej aktivity,
- burst-suppression – striedanie úsekov vysokej a nízkej aktivity,
- patologické vzory podobné epileptickým záchvatom.

Tieto vzory môžu signalizovať závažné poškodenie mozgu, ale aj určitý stupeň zotavenia, v závislosti od ich typu, trvania a časového výskytu po zástave srdca. Interpretácia týchto EEG vzorcov si však vyžaduje odborné znalosti a skúsenosti, pričom kvalitatívna analýza dlhodobých záznamov je časovo náročná a často nejednoznačná. Navyše, dostupnosť špecialistov, neurológov so špeciálnym tréningom v klinickej neurofyziológii, je v mnohých zdravotníckych zariadeniach obmedzená. Z tohto dôvodu sa v posledných rokoch zvyšuje záujem o automatizované metódy analýzy EEG, ktoré by mohli efektívne podporiť klinické rozhodovanie a prognózu pacienta.

3 Dataset

Táto práca vychádza z reálnych údajov, ktoré boli zhromaždené v rámci iniciatívy Cardiac Arrest Research (I-CARE) a poskytnuté v rámci súťaže PhysioNet Challenge 2023. Dataset obsahuje EEG záznamy a neurologické výsledky od celkovo 607 pacientov v kóme, monitorovaných v siedmich nemocniciach v Spojených štátoch a Európe (dve v Holandsku, jedna v Belgicku, tri v Bostone a jedna v New Havene).



Obr. 1: Vizualizácia datasetu.

Každý pacient bol monitorovaný pomocou 19-kanálového EEG, pričom z každého prípadu je k dispozícii viacero dátových súborov v závislosti od počtu hodín monitorovania. Výsledky pacientov boli klasifikované na základe CPC skóre do dvoch skupín: CPC 1–2 (dobrý výsledok) a CPC 3–5 (nepriaznivý výsledok).

Okrem samotných EEG záznamov a jeho CPC výsledkov, dataset poskytuje aj metadáta o pacientoch, ako napríklad:

- vek,
- pohlavie,
- informácia o návrate spontánnej cirkulácie (Return of Spontaneous Circulation – ROSC),
- výskyt zástavy srdca mimo nemocnice (Out-of-Hospital Cardiac Arrest – OHCA),
- prítomnosť komorovej fibrilácie (Ventricular Fibrillation – VF),
- použitie cielenej regulácie telesnej teploty (Targeted Temperature Management – TTM).

Samotné dáta sú poskytnuté prostredníctvom stránky iniciatívy I-CARE a sú systematicky rozdelené do priečinkov podľa jednotlivých pacientov. EEG záznamy sú uložené vo formáte súborov MATLAB (.mat) a metadáta v textovom formáte (.txt) (obr. 1).

4 Prehľad súčasného stavu riešenia

4.1 Prvé miesto na súťaži PhysioNet Challenge 2023

Jedno z úspešných riešení prezentovaných v rámci súťaže George B. Moody PhysioNet Challenge 2023 sa zameriavalo na vývoj algoritmu na predikciu obnovy vedomia pacientov po zástave srdca, a to v časovom horizonte 3 až 6 mesiacov po návrate spontánnej cirkulácie (ROSC). Riešenie vychádzalo z multimodálnych biosignálov (EEG, EKG) zozbieraných v piatich nemocniciach v USA a Európe.

Predspracovanie dát zahŕňalo zlepšenie kvality signálu pomocou pásmového a notch filtra, prevzorkovanie a normalizáciu. EEG signály boli usporiadané do 21 bipolárnych montáží a pri EKG boli štandardizované zvolené zvody. Z EEG bolo extrahovaných až 362 expertných príznakov, ktoré zahŕňali časové, spektrálne a časovo-frekvenčné metriky. Z EKG boli získané príznaky variability srdcovej frekvencie (HRV), indikátory tachyarytmií a šokovateľných rytmov.

Na klasifikáciu bol použitý prístup, ktorý kombinoval pozicionálne zakódovanie vstupov s ensemble modelom CatBoost, doplneným o viaceré typy klasifikátorov v prvej vrstve. Pre zvýšenie robustnosti a zníženie zaujatosti bol použitý systém viacerých validačných stratégií vrátane tzv. confounder-isolating cross-validation.

Navrhnutý systém dosiahol priemernú úspešnosť 0,651 $\pm$ 0,077 (v rámci 5-násobnej krížovej validácie) a na skrytej testovacej množine získal skóre 0,792, čím sa umiestnil na 1. mieste v súťaži PhysioNet Challenge 2023.

4.2 Druhé miesto na súťaži PhysioNet Challenge 2023

Ďalším výrazným riešením v rámci súťaže George B. Moody PhysioNet Challenge 2023 bol prístup tímu ComaToss, ktorý sa zameral na využitie EEG signálov na predikciu neurologickej obnovy pacientov po zástave srdca pomocou metód hlbokého učenia. Jedným z hlavných problémov bola obmedzená veľkosť EEG datasetu, čo je častá výzva v oblasti medicínskych biosignálov.

Tím tento problém riešil kombináciou predtrénovaných modelov a techník augmentácie dát. Na extrakciu príznakov z EEG signálov bol použitý hlboký model ConvNeXt, ktorý v porovnaní s inými modelmi dosiahol najlepšie výsledky. Predtrénovanie modelu na iných dátach významne zlepšilo jeho výkonnosť pri spracovaní obmedzeného EEG datasetu.

V rámci augmentácie dát boli použité techniky ako časové zrkadlenie (temporal reversal), inverzia polarity signálu (polarity inversion), CutMix na kombinovanie úsekov z viacerých signálov.

Najlepšie výsledky boli dosiahnuté práve kombináciou polarity inversion a CutMix, ktoré výrazne zvýšili robustnosť modelu voči šumu a variability v EEG dátach. Navrhnutý model dosiahol úspešnosť 0,79 na skrytej testovacej množine, čím sa tím ComaToss umiestnil na 2. mieste.

4.3 Model tímu ISIBrno-AIMT na PhysioNet Challenge 2023

Zaujímavý prístup predstavil aj tím ISIBrno-AIMT, ktorý v rámci súťaže PhysioNet Challenge 2023 navrhol hlbokú neurónovú architektúru na predikciu neurologickej obnovy pacientov po zástave srdca. Ich riešenie pozostávalo z dvojstupňového modelu: v prvej fáze boli z 5-minútových segmentov EEG signálu extrahované nízkorozmerné reprezentácie (tzv. embeddingy) pomocou hlbokého modelu, zatiaľ čo v druhej fáze sa tieto embeddingy vyhodnocovali v časovom horizonte 72 hodín pomocou Transformer enkodera. Takto navrhnutý model umožňoval zachytiť dlhodobé časové závislosti v EEG signáloch, ktoré sú kľúčové pre presnú predikciu výsledného neurologického stavu. Napriek tomu, že riešenie nebolo zaradené do oficiálneho hodnotenia v súťaži, predstavuje cenný príspevok k vývoju interpretovateľných a robustných modelov pre spracovanie EEG v akútnych medicínskych stavoch.

5 Výzvy pri predikcii dlhých časových radov

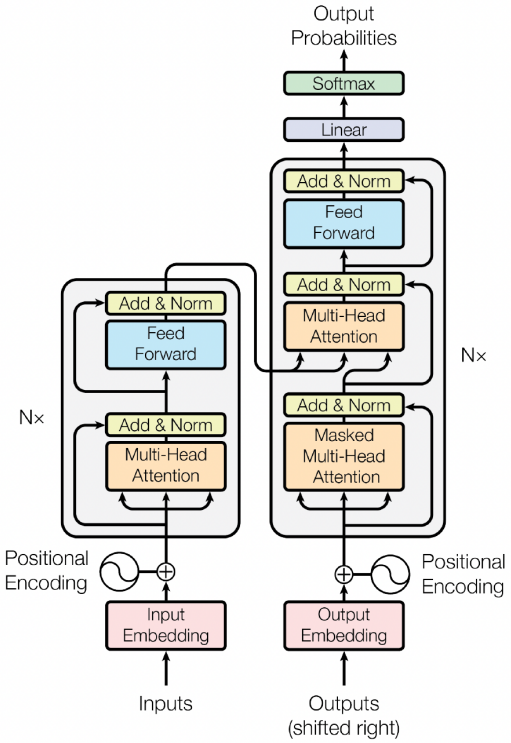
Jedným z hlavných problémov pri analýze EEG záznamov pacientov po zástave srdca je ich dĺžka a vysoká frekvencia vzorkovania. Typický EEG záznam môže trvať niekoľko hodín a zaznamenávať údaje v sekundovom alebo subsekundovom rozlíšení, čo znamená stovky tisíc až milióny časových vzoriek pre jedného pacienta. Pri dlhodobom monitorovaní sa tak vytvárajú **extrémne dlhé sekvencie**, ktoré je potrebné efektívne spracovať a využiť na predikciu klinickej prognózy. Spracovanie týchto dlhých sekvencií predstavuje niekoľko výziev:

- **Zachytenie dlhodobých závislostí:** EEG signál často obsahuje významné vzorce, ktoré sa môžu objaviť s veľkým časovým odstupom. Tradičné modely ako rekurentné neurónové siete (RNN) alebo LSTM majú tendenciu zabúdať vzdialené kontexty, čo obmedzuje ich využitie pri modelovaní dlhodobých závislostí.

- **Výpočtová a pamäťová náročnosť:** Pri použití transformerových architektúr narastá pamäťová a výpočtová zložitosť kvadraticky so vstupnou dĺžkou, čo je pri dlhodobých EEG signáloch neudržateľné. To obmedzuje možnosti aplikácie klasických transformerov na celý rozsah signálu.

- **Prítomnosť šumu a artefaktov:** Dlhodobé EEG obsahuje množstvo fyziologických aj technických rušení (napr. pohybové artefakty, zmeny kontaktu elektród), ktoré môžu znižovať kvalitu modelovania, ak sa neodfiltrujú alebo nezohľadnia.

- **Multikanálovosť a vysoká dimenzionalita:** EEG signál je zaznamenávaný pomocou viacerých (napr. 19) elektród súčasne, čím vzniká multivariačný časový rad s komplexnými priestorovo-časovými väzbami medzi kanálmi.



Obr. 2: Architektúra transformeru.

Tieto výzvy si vyžadujú špecializované modely, ktoré dokážu efektívne spracovať dlhé sekvencie bez straty kontextu, znížiť výpočtovú náročnosť a zároveň zachovať potrebnú predikčnú presnosť. Jedným z takýchto modelov je Informer, ktorý bol navrhnutý špeciálne pre úlohy dlhodobého časového predikovania a ktorý sa ukazuje ako perspektívny aj pri aplikácii na EEG dáta.

5.1 Transformery

Predtým než sa začneme zaoberať modelom Informer, je dôležité spomenúť a vysvetliť základné princípy transformerov. Transformery (obr. 2) sú dnes jednou z najvýkonnejších a najflexibilnejších architektúr pre spracovanie sekvenčných dát, čo potvrdzuje ich široké použitie v rôznych oblastiach od spracovania prirodzeného jazyka (NLP), cez počítačové videnie až po časové rady a bio-signály.

Dôvod vysokého úspechu transformerov spočíva predovšetkým v ich schopnosti paralelne spracovávať celú vstupnú sekvenciu a efektívne modelovať dlhodobé závislosti medzi jej prvkami pomocou mechanizmu self-attention. Na rozdiel od tradičných rekurentných neurónových sietí (RNN), ktoré spracúvajú vstup po jednom prvku a často zápasia s problémom miznúcich gradientov, transformery vyhodnocujú vzťahy medzi všetkými prvkami sekvencie naraz.

5.1.1 Princíp self-attention a Q, K, V na jednoduchom príklade

Predstavme si vetu so slovami: *"mačka", "vidí", "myš"*. Pre každé slovo sa vytvoria tri vektory: *query* (Q), *key* (K) a *value* (V). Napríklad, pre slovo „mačka“ ako query spočítame skalárne súčiny (dot-product) jeho Q vektora so všetkými K vektormi slov v sekvencii (*"mačka", "vidí", "myš"*).

Výsledné skóre následne normalizujeme pomocou funkcie Softmax, čím získame váhy určujúce, koľko pozornosti „mačka“ venuje jednotlivým slovám.

Tieto váhy sa použijú na vážený súčet hodnôt V, čo vedie k novej reprezentácii slova „mačka“ so zapracovaným kontextom z ostatných slov.

5.1.2 Multi-head attention

Zatiaľ čo základná self-attention mechanika umožňuje modelu zachytiť vzťahy medzi jednotlivými pozíciami v sekvencii, mechanizmus multi-head attention (obr. 3) ide ešte ďalej a umožňuje modelu pozerať sa na dáta z viacerých perspektív súčasne.

Každá hlava (head) vo vnútri multi-head attention predstavuje vlastnú verziu self-attention, ktorá má svoje vlastné váhové matice pre Q, K a V. Tieto jednotlivé hlavy sa učia zachytávať rôzne druhy závislostí, napríklad jedna hlava sa môže zamerať na krátkodobé vzťahy, iná na dlhodobé, ďalšia na gramatické súvislosti atď.

Formálne je výpočet multi-head attention nasledovný:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{1}$$

kde každá hlava je definovaná ako:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

Každá sada váhových matíc W_i^Q, W_i^K, W_i^V je špecifická pre konkrétnu hlavu i, a výstupy všetkých hláv sa následne zreťazia (concat) a transformujú pomocou výstupnej projekčnej matice W^O.

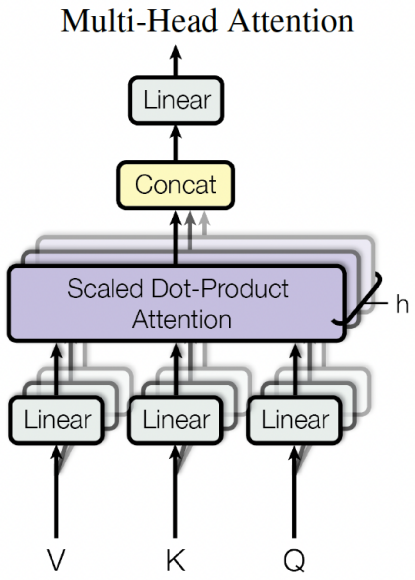
Táto architektúra umožňuje modelu zachytiť bohatšiu reprezentáciu kontextu, pretože rôzne hlavy sa špecializujú na rôzne typy informácií.

Príklad: Ak analyzujeme vetu *”mačka vidí myš”*, jedna hlava sa môže sústrediť na syntaktické závislosti (napr. kto je subjekt a objekt), zatiaľ čo iná môže sledovať časový alebo tematický kontext. Vďaka paralelnému spracovaniu všetkých týchto aspektov dokáže transformer efektívne porozumieť komplexnému významu textu alebo iných dát.

5.1.3 Rôzne implementácie pre rôzne typy dát

Transformery sa dokážu flexibilne prispôsobiť rôznym dátovým typom a dimenziám:

- V NLP pracujú so slovnými tokenmi a zachytávajú vzťahy medzi slovami v texte.
- V počítačovom videní rozdeľujú obrázky na menšie bloky (patches) a spracovávajú ich ako sekvenciu.



Obr. 3: Vizualizácia multi-head attention mechanizmu.

- Pri časových radoch dokážu modelovať dlhodobé závislosti v dátach, napríklad pri analýze EEG alebo senzorových dát.

Vďaka tomu sú transformery univerzálnym nástrojom na riešenie komplexných problémov v rôznych oblastiach.

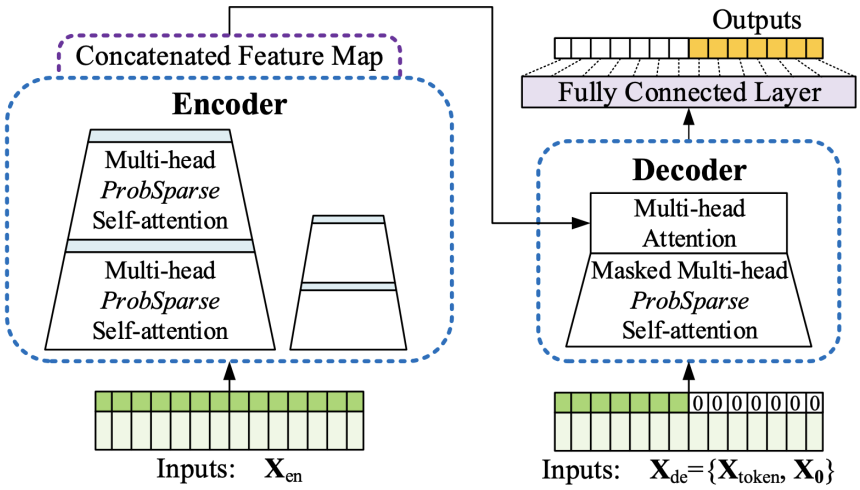
5.1.4 Príklad interpretácie Q, K, V pri EEG signáloch

V prípade EEG signálov, ktoré sú vysoko časovo závislé a často multikanálové, predstavujú vstupy do mechanizmu self-attention rôzne časti signálu v čase. Typicky sa EEG signál rozdelí na menšie časové segmenty (napr. 0.5-sekundové úseky), ktoré sa následne považujú za vstupné tokeny.

Pre každý z týchto segmentov sa následne lineárnymi transformáciami vygenerujú tri rôzne reprezentácie:

- **Query (Q)**: „Na ktorý segment sa pozerám?“ – reprezentuje aktuálny časový segment, pre ktorý chceme získať kontext;
- **Key (K)**: „Čo je v ostatných segmentoch?“ – reprezentuje všetky ostatné segmenty v sekvencii, s ktorými sa query porovnáva;
- **Value (V)**: „Čo si z týchto segmentov chcem zapamätať?“ – nesie informáciu, ktorá sa vážene agreguje podľa podobnosti medzi query a key.

Mechanizmus self-attention tak umožňuje zohľadniť kontext celého signálu pri spracovaní každého jednotlivého segmentu. Napríklad, aktivita na čele (frontálne elektródy) môže byť informatívna pri rozhodovaní o signáli, ktorý vznikol o pár sekúnd neskôr v týlnej oblasti (okcipitálne elektródy).



Obr. 4: Architektúra modelu Informer.

Príklad: Predstavme si EEG signál s 16 kanálmi, dĺžkou 10 sekúnd a vzorkovacou frekvenciou 256 Hz, teda máme 2560 vzoriek. Rozdelením na segmenty po 128 vzorkách (0.5 sekundy) získame 20 segmentov. Každý z nich sa spracuje (napr. pomocou CNN) na vektor dĺžky d_{model}, a ďalej sa použije ako token v transformeri.

5.2 Model Informer

Pre úlohu predikcie alebo klasifikácie nad dlhými EEG sekvenciami je nevyhnutné zvoliť model, ktorý dokáže efektívne zachytávať dlhodobé závislosti. Tradičné modely ako rekurentné neurónové siete (RNN) alebo ich vylepšené varianty ako LSTM (Long Short-Term Memory) majú síce určitú schopnosť pracovať so sekvenčnými dátami, no pri veľmi dlhých časových radoch strácajú schopnosť udržať informáciu z dávnejších častí signálu. To je spôsobené problémom miznúcich gradientov, obmedzenou kapacitou pamäte a praktickými obmedzeniami výpočtových prostriedkov.

Z tohto dôvodu sa v posledných rokoch do popredia dostávajú modely založené na architektúre transformerov, ktoré sú schopné efektívnejšie modelovať globálne vzťahy medzi prvkami v sekvencii. Avšak klasické transformery majú kvadratickú časovú a pamäťovú zložitosť vzhľadom na dĺžku vstupu, čo znemožňuje ich použitie na extrémne dlhé sekvencie.

Model Informer (obr. 4), predstavený v práci *Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting* (Zhou et al., 2021), predstavuje riešenie týchto problémov. Bol navrhnutý špecificky pre predikciu v dlhých časových radoch (LSTF – Long Sequence Time-Series Forecasting). Disponuje tromi kľúčovými inováciami:

- **ProbSparse Attention** – upravený mechanizmus pozornosti, ktorý výberovo uprednostňuje iba tie pozorovania, ktoré významne prispievajú k výstupu. Tým znižuje časovú a pamäťovú náročnosť na $O(L \log L)$.
- **Self-Attention Distilling** – technika, ktorá postupne redukuje veľkosť vstupnej sekvencie

medzi vrstvami modelu. Cieľom je extrahovať dominantné črty signálu a zároveň zabrániť zahlteniu nepodstatnými detailmi.

- **Generatívny dekodér** – namiesto klasického autoregresívneho dekódovania po jednom kroku naraz, Informer dekóduje celé sekvencie naraz. To výrazne zvyšuje rýchlosť inferencie pri dlhých predikciách.

Tieto vlastnosti robia z Informera výnimočný model pre spracovanie údajov ako sú EEG záznamy s vysokým časovým rozlíšením a dlhým trvaním. Vďaka efektívnej implementácii pozornosti a schopnosti zachytiť globálne vzťahy je vhodný aj pre úlohy klasifikácie neurologického stavu pacientov na základe dlhodobých signálov.

Model Informer je open-source a je dostupný na platforme GitHub spolu s množstvom predpripravených datasetov a nástrojov, čo uľahčuje jeho aplikáciu aj v medicínskom výskume.

5.2.1 ProbSparse Self-Attention

Jedným z hlavných problémov klasického mechanizmu *self-attention*, ktorý je základom transformerových modelov, je jeho kvadratická časová a pamäťová zložitosť vzhľadom na dĺžku vstupu L, konkrétne $O(L^2)$.

Model Informer preto zavádza ProbSparse Self-Attention – efektívnu aproximáciu self-attention, ktorá znižuje výpočtovú náročnosť na $O(L \log L)$, pričom si zachováva schopnosť modelovať globálne závislosti.

Základný vzorec klasickej self-attention:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{3}$$

Kde $Q \in R^{L\times d}$ sú *queries*, $K \in R^{L\times d}$ sú *keys*, $V \in R^{L\times d_v}$ sú *values*, a d je rozmer modelu.

ProbSparse modifikácia: Namiesto výpočtu všetkých skalárnych súčinov (dot-produktov) medzi všetkými vektormi *query* a *key*, ktoré určujú mieru pozornosti, sa v mechanizme ProbSparse počítajú iba tie najdôležitejšie. Ich výber sa riadi podľa tzv. miery významnosti, ktorá je definovaná nasledovne:

$$M(q_i, K) = \max_j \left(\frac{q_i k_j^\top}{\sqrt{d}}\right) - \frac{1}{L_K}\sum_{j=1}^{L_K}\left(\frac{q_i k_j^\top}{\sqrt{d}}\right) \tag{4}$$

Tento vzorec (tzv. max-mean sparsity measurement) meria rozdiel medzi najväčšou hodnotou pozornosti a priemernou pozornosťou daného query q_i na všetky keys. Čím väčší tento rozdiel je, tým dôležitejší je daný query pre výstup.

Na základe tejto metriky sa vyberie len $u = c \cdot \ln L_Q$ najdôležitejších queries pre výpočet pozornosti, kde c je konštanta. To znamená, že namiesto výpočtu $O(L^2)$ dot-produktov sa počíta

Algorithm 1 ProbSparse self-attention

Require: Tensor $\mathbf{Q} \in \mathbb{R}^{m \times d}$, $\mathbf{K} \in \mathbb{R}^{n \times d}$, $\mathbf{V} \in \mathbb{R}^{n \times d}$

1: **print** set hyperparameter c, $u = c \ln m$ and $U = m \ln n$
2: randomly select U dot-product pairs from $\mathbf{K}$ as $\bar{\mathbf{K}}$
3: set the sample score $\bar{\mathbf{S}} = \mathbf{Q}\bar{\mathbf{K}}^\top$
4: compute the measurement $M = \max(\bar{\mathbf{S}}) - \text{mean}(\bar{\mathbf{S}})$ by row
5: set Top-u queries under M as $\bar{\mathbf{Q}}$
6: set $\mathbf{S}_1 = \text{softmax}(\bar{\mathbf{Q}}\mathbf{K}^\top/\sqrt{d}) \cdot \mathbf{V}$
7: set $\mathbf{S}_0 = \text{mean}(\mathbf{V})$
8: set $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_0\}$ by their original rows accordingly

Ensure: self-attention feature map $\mathbf{S}$.

Obr. 5: Algoritmus ProbSparse Self-Attention mechanizmu.

len $O(L \ln L)$, čo výrazne znižuje čas aj pamäť.

$$\text{ProbSparseAttention}(Q, K, V) = \text{Softmax}\left(\frac{Q^* K^\top}{\sqrt{d}}\right) V \tag{5}$$

Kde Q^* obsahuje len tie queries, ktoré boli vybrané na základe sparsity miery $M(q_i, K)$.

Popis jednotlivých krokov ProbSparse Self-Attention algoritmu (obr. 5):

1. **Nastavenie hyperparametrov:** Zvolí sa parameter c, na základe ktorého sa vypočíta počet najdôležitejších queries $u = c \cdot \ln m$ a počet náhodne vybraných dot-product párov $U = m \cdot \ln n$, kde m je počet queries a n je počet keys.

2. **Náhodný výber:** Z množiny K sa náhodne vyberie U keys a vytvorí sa ich podmnožina $\bar{K}$, čím sa redukuje výpočtová náročnosť.

3. **Výpočet skóre:** Spočítajú sa dot-produkty medzi všetkými queries Q a vybranými keys $\bar{K}$, čím vznikne matica skóre $\bar{S} = Q\bar{K}^\top$.

4. **Výpočet sparsity miery:** Pre každý query sa spočíta hodnota $M = \max(\bar{S}) - \text{mean}(\bar{S})$, teda rozdiel medzi maximálnym a priemerným skóre v riadku. Táto max-mean metrika je numericky stabilná a menej citlivá na nulové hodnoty. V praxi ide o jednoduchý spôsob, ako kvantifikovať dôležitosť daného query – čím väčší rozdiel, tým viac sa query viaže len na niekoľko silných keys, a teda je informatívnejší.

   Tento krok však nie je fixne daný a je možné ho upraviť podľa potreby aplikácie. Namiesto aritmetického priemeru možno použiť aj medián, čím sa dosiahne vyššia odolnosť voči odľahlým hodnotám. Podobne je možné experimentovať aj so zložitejšími mierami „koncentrovanosti“ pozornosti, ako napríklad rozptyl dot-produktov, entropia normalizovaných skóre, alebo priemer z top-k hodnôt namiesto maxima. V niektorých implementáciách je dokonca možné nahradiť túto mieru malou neurónovou sieťou, ktorá sa učí priamo počas

tréningu modelu.

5. **Výber najdôležitejších queries:** Vyberie sa top-u queries s najvyššími hodnotami sparsity miery a označia sa ako $\bar{Q}$.

6. **Výpočet pozornosti pre $\bar{Q}$:** Pre vybrané queries $\bar{Q}$ sa spočíta klasická self-attention: $\text{Softmax}(\bar{Q}K^\top/\sqrt{d}) \cdot V$.

7. **Odhad pre zvyšné queries:** Queries, ktoré neboli zaradené do $\bar{Q}$, neprechádzajú výpočtom pozornosti. Namiesto toho im je priradený jednoduchý priemer všetkých hodnôt V ako rýchly, ale použiteľný odhad.

8. **Zlúčenie výsledkov:** Výstupy sa nakoniec zoradia podľa pôvodného poradia queries, čím vznikne finálna feature mapa self-attention mechanizmu.

5.2.2 Self-attention distilling

Enkóder modelu Informer bol navrhnutý tak, aby zvládal spracovanie extrémne dlhých sekvencií pod obmedzením pamäťových nárokov. Každý vstup X^t je po úvodnej reprezentácii prevedený na maticu $X^t_{en} \in R^{L \times d_{model}}$, kde L je dĺžka vstupnej sekvencie a d_{model} je dimenzionalita skrytého priestoru.

Jednou z kľúčových techník v enkóderi je tzv. self-attention distilling, ktorá prirodzene nadväzuje na využitie ProbSparse pozornosti. Keďže kombinácií hodnôt V v pozornosti je veľa a mnohé sú redundantné, cieľom distillácie je zachovať len dominantné črty sekvencie a postupne znižovať časovú dimenziu výstupu.

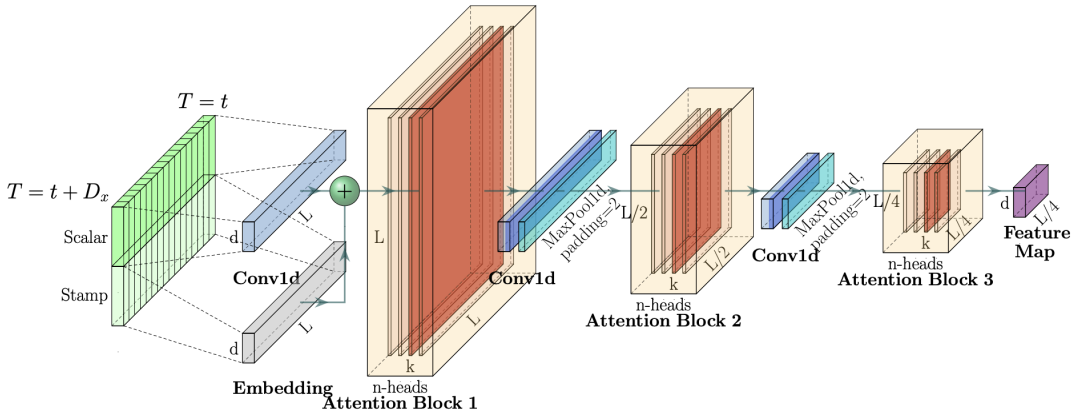
Tento mechanizmus je inšpirovaný dilatovanými konvolúciami. Na výstup každej vrstvy pozornosti sa aplikuje 1D konvolučný filter so šírkou jadra 3, následne aktivačná funkcia ELU a nakoniec max-poolovanie s posunom 2, ktoré redukuje dĺžku výstupu na polovicu. Tento proces je formálne definovaný ako:

$$X^t_{j+1} = \text{MaxPool}\left(\text{ELU}\left(\text{Conv1d}\left(\left[X^t_j\right]_{AB}\right)\right)\right),$$

kde $[X^t_j]_{AB}$ reprezentuje výstup z attention bloku v j-tej vrstve (vrátane ProbSparse pozornosti a vrstiev normovania a reziduálneho spojenia).

Aby sa zvýšila robustnosť tohto procesu, Informer stavia viacero paralelných zásobníkov (stacks), kde každý z nich pracuje s inak veľkou časťou vstupu (napr. polovica, štvrtina...). Výstupy všetkých zásobníkov sú nakoniec spojené do jedného výstupu enkódera.

Tento prístup výrazne znižuje pamäťovú zložitosť na $\mathcal{O}((2-\varepsilon)L\log L)$ a zároveň umožňuje extrakciu globálnych závislostí v dátach s veľmi dlhým trvaním, ako sú EEG záznamy.



Obr. 6: Architektúra jedného zásobníka enkódera modelu Informer. Prvý zásobník spracováva celú vstupnú sekvenciu, druhý iba polovicu, atď. Červené bloky predstavujú maticové výpočty pozornosti, ktoré sa kaskádovito zmenšujú vďaka mechanizmu *self-attention distilling*. Výstupy zo všetkých zásobníkov sú na konci spojené.

5.2.3 Generatívny dekóder

Dekóder v modeli Informer je navrhnutý na generovanie dlhých sekvenčných výstupov efektívnym spôsobom. Skladá sa z dvoch vrstiev multi-head pozornosti a využíva modifikovanú verziu ProbSparse self-attention spolu s masked attention, ktorá zabezpečuje, že každá pozícia vo výstupe má prístup iba k predchádzajúcim (a nie budúcim) časovým krokom.

Na rozdiel od tradičného autoregresívneho dekódovania, kde sa výstup generuje po jednotlivých krokoch, Informer využíva tzv. generative inference, teda predikciu celej výstupnej sekvencie v jednom priamom prechode modelom. Vstupom do dekodéra je start token, t. j. krátka známa časť sekvencie, spolu s nulovým placeholderom pre cieľovú sekvenciu. Tento prístup výrazne urýchľuje inferenciu, najmä pri dlhých predikčných oknách.

V tejto práci sa však zameriavame na klasifikačný problém, nie na predikciu budúcich hodnôt. Preto dekóderovú časť modelu Informer nevyužívame a pracujeme výlučne s enkóderom, ktorý slúži na extrakciu reprezentácií z dlhých EEG sekvencií.

6 Prehľad splnených úloh

Doteraz vykonané kroky v rámci tejto práce zahŕňajú teoretickú prípravu, analýzu dát a predprípravu na experimentálnu časť.

- **Spracovanie relevantných vedeckých prác:** Napr. článok *Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting*, ktorý získal ocenenie Best Paper Award na celosvetovej konferencii o umelej inteligencii v roku 2021.
- **Predspracovanie datasetu:** V spolupráci so spoločnosťou VSL Software bol pôvodný rozsiahly dataset EEG signálov (veľkosti približne 1,5 TB) redukovaný na cca 130 GB, pričom boli vykonané nasledovné kroky:

- odstránené boli časové rady s konštantným priebehom, ktoré neobsahovali žiadnu informatívnu zložku,
- vyradené boli tiež sekvencie s dlhým výpadkom signálu alebo iným typom chybových údajov,
- boli zachované len kvalitné a použiteľné signály vhodné na následné trénovanie modelov.

Týmto krokom bola vytvorená kvalitná báza pre následnú implementáciu a experimentálne overenie výkonnosti modelov.

7 Najbližšie kroky

V ďalšej fáze práce bude hlavným cieľom implementácia modelu Informer a jeho aplikácia na predspracované EEG dáta. Model bude testovaný v rôznych konfiguráciách s cieľom overiť jeho schopnosť efektívne modelovať dlhodobé závislosti v sekvenčných signáloch.

Okrem základnej implementácie sa plánuje vytvorenie rozšíreného modelu, ktorý bude navyše pracovať s dostupnými metadátami o pacientoch (napr. vek, pohlavie), pokiaľ sa ukáže, že tieto doplnkové informácie vedú k zvýšeniu presnosti predikcie.

Súčasťou ďalšieho postupu bude aj porovnanie Informera s inými modelmi, či už z oblasti transformerových architektúr alebo hlbokých neurónových sietí (napr. LSTM, CNN). Hlavným cieľom bude preskúmať ich výkonnosť z pohľadu časovej zložitosti, presnosti výsledkov a celkovej optimality riešenia pre veľké sekvencie EEG dát.

- **Plánované použité metriky:**
  - **MSE (Mean Squared Error)** — základná metrika pre regresné úlohy.
  - **MAE (Mean Absolute Error)** — doplnková metrika, menej citlivá na extrémy.
  - **Inference time** — priemerný čas potrebný na spracovanie jednej sekvencie.
  - **Počet parametrov a pamäťová náročnosť modelu**.
- **Experimentovanie s dĺžkami vstupnej/výstupnej sekvencie:** s cieľom nájsť optimálny pomer medzi rozsahom informácie a výpočtovou náročnosťou.
- **Vizualizácia pozornostı́:** analýza mechanizmu attention s cieľom pochopiť, ktoré časti EEG signálu model považuje za relevantné.

Literatúra

[1] AMORIM, Edilberto, et al., 2023. The international cardiac arrest research consortium electroencephalography database. In: *Critical Care Medicine*. Vol. 51, no. 12, p. 1802-1811. DOI: 10.1101/2023.08.28.23294672v1

[2] ZABIHI, Morteza, et al., 2023. Hyperensemble learning from multimodal biosignals to robustly predict functional outcome after cardiac arrest. In: *Computing in Cardiology (CinC). IEEE*. p. 1-4. DOI: 10.22489/CinC.2023.142

[3] KIM, Dong-Kyu, et al., 2023. Predicting Neurological Outcome After Cardiac Arrest Using a Pretrained Model with Electroencephalography Augmentation. In: *Computing in Cardiology (CinC). IEEE*. p. 1-4. DOI: 10.22489/CinC.2023.077

[4] PAVLUS, Jan, et al., 2023. Using Embedding Extractor and Transformer Encoder for Predicting Neurological Recovery from Coma After Cardiac Arrest. In: *Computing in Cardiology (CinC). IEEE*. p. 1-4. DOI: 10.22489/CinC.2023.054

[5] VASWANI, Ashish, et al., 2017. Attention is all you need. In: *Advances in neural information processing systems*. vol. 30. DOI: arXiv:1706.03762v7

[6] ZHOU, Haoyi, et al., 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 35, no. 12, p. 11106-11115. DOI: https://doi.org/10.1609/aaai.v35i12.17325