

Budapesti Műszaki és Gazdaságtudományi Egyetem

Villamosmérnöki és Informatikai Kar

Dancsó Marcell

Jelnyelv fordítás

TDK dolgozat

Konzulens

BUDAPEST, 2023

Tartalomjegyzék

[Összefoglaló 4](#_Toc149434889)

[Abstract 6](#_Toc149434890)

[1 Bevezetés 8](#_Toc149434891)

[1.1 A dolgozat felépítése 8](#_Toc149434892)

[1.2 Jelnyelv alapok 9](#_Toc149434893)

[1.3 Amerikai jelnyelv (ASL) 10](#_Toc149434894)

[1.3.1 Ujjbetűzés az amerikai jelnyelvben 10](#_Toc149434895)

[1.3.2 Általános ASL 11](#_Toc149434896)

[2 Kapcsolódó kutatások 12](#_Toc149434897)

[2.1 Hagyományos képfeldolgozáson alapuló módszerek 12](#_Toc149434898)

[2.1.1 Jelnyelv automatikus fordítása többfolyamos 3D CNN-nel és mesterséges mélységtérképek generálása 12](#_Toc149434899)

[2.2 Segédeszközt használó megoldások 13](#_Toc149434900)

[2.2.1 Kesztyűt használó kutatások 13](#_Toc149434901)

[2.2.2 SignRing: Amerikai jelnyelv-felismerés IMU szenzorral ellátott gyűrűkkel 14](#_Toc149434902)

[2.3 A dolgozat célja 14](#_Toc149434903)

[3 Megközelítés 16](#_Toc149434904)

[3.1 Póz felismerés 16](#_Toc149434905)

[3.1.1 Open pose 17](#_Toc149434906)

[3.1.2 Mediapipe holistic és hands 18](#_Toc149434907)

[4 Statikus ujjbetűzés 20](#_Toc149434908)

[4.1 Adathalmaz 20](#_Toc149434909)

[4.1.1 Interpolációs algoritmusok 21](#_Toc149434910)

[4.1.2 Fekete fehér képek színezése mély tanuláson alapuló modellekkel 23](#_Toc149434911)

[4.1.3 Minőség javító algoritmusok összehasonlítása 24](#_Toc149434912)

[4.1.4 Saját ASL adathalmaz 25](#_Toc149434913)

[4.2 Modellek 26](#_Toc149434914)

[4.3 Kiegészítő algoritmusok 27](#_Toc149434915)

[4.3.1 Kimenet korrigálása valószínűségi alapon 27](#_Toc149434916)

[4.3.2 Nem szándékos mozgások elkülönítése 28](#_Toc149434917)

[4.3.3 Hibák korrekciója nagy nyelvi modellekkel 29](#_Toc149434918)

[4.4 Értékelés 30](#_Toc149434919)

[5 Ujjbetűzés szekvenciális bemenetből 31](#_Toc149434920)

[5.1 Adathalmaz 31](#_Toc149434921)

[5.2 Modellek 32](#_Toc149434922)

[5.2.1 Encoder-Decoder architektúra 32](#_Toc149434923)

[5.2.2 Saját modell kialakítása 33](#_Toc149434924)

[5.2.3 Tanítás 36](#_Toc149434925)

[5.3 Kiegészítő algoritmusok 37](#_Toc149434926)

[5.4 Értékelés 37](#_Toc149434927)

[6 Hagyomás jelelés 38](#_Toc149434928)

[7 Általános jelnyelv fordítása 39](#_Toc149434929)

[8 Köszönetnyilvánítás 40](#_Toc149434930)

[9 Irodalomjegyzék 41](#_Toc149434931)

[Irodalomjegyzék 42](#_Toc149434932)

[Függelék 43](#_Toc149434933)

Összefoglaló

A nyelvfeldolgozás terén elért sikerek rengetek kaput nyitottak ki előttünk. A hangalapú asszisztensek, alap esetben is hasznos, de legtöbbször forradalmi funkciókat hoznak életünkbe. A mesterséges intelligencián alapuló „voice to text” modellek manapság gyakorlatilag tetszőleges nyelvről képesek felismerni szöveget, valamint a közelmúltban nyelvfeldolgozás területén elért eredményeknek köszönhetően pedig nincsenek határok, a felhasználók, és fejlesztők előtt. Illetve előbbi állítás sajnos csak egy nagy ferdítéssel igaz. Ugyanis világszerte körülbelül 1,5 milliárd ember hallássérült, és több mint 70 millió ember használja a jelnyelvet, mint elsődleges kommunikációs formát. Számukra „saját nyelvükön” ezek a funkciók nem, vagy csak korlátozottan érhetőek el. A jelnyelv egy speciális fajtája az ujjbetűzés. Gyakran használják nevek, címek, telefonszámok, valamint olyan fogalmak közvetítése során, melyekre nincs bevett gesztus. Egy tapasztalt jelelő képes közel kétszer olyan sebesen betűzni, mint egy virtuális billentyűzeten pötyögő egyén, nem beszélve arról ha minden jelet alkalmazva kommunikál. Így égető egy számukra is kényelmesen használható interfész kialakítása.

Még jelentősebb problémát vet fel a kommunikáció kérdése. Manapság a nyelvek közötti gépi fordítás széles körben és formátumban elérhető az internet kapcsolattal rendelkezőknek. Régen túl vagyunk már az egyszerű szótár alapú fordítókon, a különböző transzformer, és nagy nyelvi modellek térhódításával gyakorlatilag tetszőleges nyelvek között megoldható a kommunikáció. Ez alól kivétel azonban a jelnyelv. Nem létezik olyan megoldás a piacon, ami segíti a jelelő kommunikációját a jelelni nem tudó személy felé. A terület előrehaladásán nem segít, hogy a jelnyelv csakúgy mint a verbális párjai, nem standardizált. Szinte minden nemzetnek saját jelnyelve van, nem beszélve a helyi sajátosságokról, és dialektusokról. Tovább nehezíti a feladatot, hogy a rendelkezésre álló, nagy méretű, és jó minőségű adatbázisok száma eltörpül a klasszikus nyelvfeldolgozásban megszokottaktól.

Munkámmal a széles tömegek számára elérhető jelnyelv fordító rendszerek fejlesztéséhez járulok hozzá. Megvizsgálom többek között a kesztyűvel történő felismerés, valamint hagyományos képfeldolgozás eredményeit, továbbá részletesen foglalkozom a póz approximáción alapuló algoritmusokkal. A nyílt kérdés utóbbival kapcsolatban, hogy bár rendkívül jó arányban tömöríti a képi adatot, ami lehetővé tenné nagy mennyiségű, egységes adatbázis konstruálását, de vajon tart-e ott a technológia, hogy megbízható módon kódoljon minden fordításhoz szükséges információt?

Az amerikai jelnyelven belül külön-külön vizsgálom az ujjbetűzés, és általános jelbeszéd lehetőségeit, kitérve a pillanatképből, valamint mozgásszekvenciából dolgozó megoldásokra. A teljesség igénye nélkül összehasonlítom a feladatra adaptált konvolúciós, rekurrens, LSTM, Transzformer háló architektúrákat. A jelelés nyelvtani adottságai következtében, hiányoznak a segédigék, valamint gyakran más a szórend a hagyományos angolhoz képest. A primitív fordítás eredményeit ezért kontextus függő módon transzformálom generatív nagy nyelvi modellekkel, valamint az ujjbetűzés esetében a megbízhatóság növelésére is felhasználom őket.

A dolgozatomban bemutatom az adatgyűjtés és a tervezés lépéseit, az alkalmazott mesterséges intelligencia algoritmusok részleteit, valamint elemzem a fejlesztés során felmerült tervezői döntéseket, és alternatíváikat. Ezen felül bemutatásra kerül az általam készített teljes megoldás, mely nagymértékben segíthet a hallássérültekkel való kommunikáció során.

Abstract

The successes achieved in the field of natural language processing have opened up numerous doors before us. Voice-based assistants are not only useful, but they often introduce revolutionary functionalities into our lives. Nowadays, artificial intelligence-based voice-to-text models are capable of recognizing text from virtually any language, and thanks to recent advancements in natural language processing, there are no boundaries for users and developers. However, the previous statement is a significant distortion of the truth. Globally, approximately 1.5 billion people are hearing impaired, and more than 70 million people use sign language as their primary means of communication. For them, these functionalities are either not available in their „own language”, or only in a limited manner. Sign language, particularly finger-spelling, is a special form of communication often used for communicating names, addresses, and phone numbers, or conveying concepts without standard gestures. An experienced signer can spell almost twice as fast as an individual typing on a virtual keyboard, let alone when using all available signs for communication. Hence, there is an urgent need to create an interface that is comfortable for them to use.

An even more significant issue arises in the realm of communication. Machine translation between spoken languages is widely available to those with internet access in various formats. We have long surpassed simple dictionary-based translators, and with the rise of different transformers and large language models, communication between virtually any language has long become a reality. However, sign language even today remains an exception. No solution on the market assists signers in communicating with non-signing individuals. The progress in this area is hindered by the fact that sign language, like its verbal counterparts, is not standardized. Almost every nation has its own sign language, not to mention regional variations and dialects. Furthermore, the task is complicated by the fact that the available large and high-quality datasets are minuscule compared to what is customary in classical natural language processing.

My work contributes to the development of sign language translation systems accessible to the broader masses. I have conducted research encompassing various aspects, including glove-based recognition and traditional image processing methods. However, my primary focus was directed towards the comprehensive exploration of using pose approximation algorithms. While the latter is highly promising as it compresses visual data effectively, enabling the construction of large, homogeneous databases, the question of whether current technology is capable of reliably encoding all the information needed for translation has so far remained open.

Within American Sign Language, I separately explored finger-spelling and general sign language possibilities, paying close attention to solutions working from both snapshots and motion sequences. Without aiming for completeness, I compared convolutional, recurrent, LSTM, and Transformer network architectures adapted for the task. Due to the grammatical characteristics of sign language, auxiliary verbs are missing, and the word order is often different from traditional English when translated word by word. Therefore, I processed the results contextually using generative large language models and employed them to enhance the quality and reliability of the translation.

In my thesis, I walk through the steps of data collection, provide details about the applied artificial intelligence algorithms, and analyze the design decisions and their alternatives that emerged during development. Furthermore, I present the comprehensive solution I've developed, offering significant support for communication with the hearing impaired.

# Bevezetés

A nyelvfeldolgozás területén hatalmas áttöréseket értünk el az elmúlt években. A mai „voice to text” modellekkel lehetőségünk van rendkívül sokszínű és természetes módon interakcióba lépni az eszközeinkkel. Kiegészülve a világot alapjaiban megrengető intelligens nagy nyelvi modellekkel, saját virtuális asszisztenst alkalmazhatunk, chat formában érhetjük el a világ tudását és nem utolsó sorban kinyílik a kapu a szabad kommunikáció előtt tetszőleges nyelvek között. Az hasonló rendszerek egyik sajnálatos velejárója, hogy csupán hagyományos nyelveken érhetőek el. Bár látszólag a technikai tudás rendelkezésre áll, még sem jelent meg még számottevő jelnyelvet támogató rendszer a piacon. A jelnyelv szöveggé formálásával közel 70 millió embernek nyílna lehetősége saját nyelvén interakcióba lépni okos eszközökkel, nem beszélve a jelelni nem tudók felé irányuló kommunikációról. Nagy segítséget nyújthatna továbbá a tanulásban azoknak, akik annak ellenére, hogy hallássérüléssel kénytelenek élni, nincs lehetőségük megtanulni, mind pénzügyi, mind tanulást segítő eszközök híján. Ezen csoport mérete meglepően még jelentősebb, mint a jelelni tudók száma, bizonyítva a nyelv komplexitását, és ezzel a feladat nehézségét.

## A dolgozat felépítése

A bevezető után, rövid leírást adok a jelnyelvek felépítéséről, illetve eltéréseiről a hagyományos nyelvekhez képest, részletesebben foglalkozva az amerikai jelnyelv sajátosságaival.

Megvizsgálom a téma irodalmát, kitérve a segédeszközöket felhasználó, valamint hagyományos képfeldolgozáson alapuló algoritmusokra. Összehasonlítom más kutatások eredményeit, valamint az alkalmazott módszerek hátrányait. Utóbbiból kiindulva szemléltetem a dolgozat motivációját, célkitűzéseit.

Ezt követően nagy léptékben haladva, ismertetem a kutatás menetét, alkalmazott módszertanokat, valamint a kiindulási alapként szolgáló megoldások algoritmikus hátterét.

A következő fejezetekben külön-külön vizsgálom a póz approximáción alapuló ujjbetűzés lehetőségeit mind pillanatképekből, valamint szekvenciális képkockákból, a hagyományos jelelésre kialakított megoldásom, valamint részletezem a kettő ötvözésének lehetőségét. Részletes leírást adok az adatgyűjtés folyamatáról, egyes adathalmazok leírásáról, illetve feldolgozásukról. Ismertetem a kialakított modelleket, nagy hangsúlyt fektetve a tervezői döntések indoklására, alternatívák feltárására, továbbá számos metrika szerint kiértékelem, és összehasonlítom őket. A legjobb modellekhez különböző „usecase”-ek szerint csoportosítva kiegészítő algoritmusokat tervezek.

Végül összefoglalom az eredményeket, további javaslatokat teszek a fejlesztésre, és megválaszolom a kérdést: Vajon a póz approximáció alkalmas-e jelen formájában komplex rendszerek kialakítására.

## Jelnyelv alapok

A jelnyelv egy vizuális gesztusokra nyelv, amelyet a hallássérült és a halló közösségek egyaránt használnak a kommunikációhoz. A beszélt nyelvektől eltérően a jelnyelvben a jelentést a kézmozdulatok, az arc kifejezései és a testtartás hordozzák.

Míg sokan tévesen úgy gondolják, hogy a jelnyelv egyetemes, valójában számos különböző jelnyelv létezik világszerte. Minden ország saját jelnyelvvel rendelkezik. Érdekes például, hogy az amerikai jelnyelv (ASL) és a brit jelnyelv (BSL) annyira különbözőek, hogy nem érthetőek egymás számára, annak ellenére, hogy mindkét ország angolul beszél. A két nyelvnek mellesleg teljesen különböző fejlődése, míg az ASL-nek francia jelnyelvi gyökerei vannak, addig a BSL teljesen külön fejlődött. Továbbá hasonlóan a verbális nyelvekhez, itt is léteznek dialektusok, helyi sajátosságok, ami további komplexitáshoz vezet intelligens megoldások kialakításakor.

A legtöbb jelnyelv egy speciális részhalmaza az ujjbetűzés, amely kézzel formázott betűk sorozata. Néhány jelnyelvben az ujjbetűzést gyakran használják nevek, idegen szavak vagy specifikus terminológiák kifejezésére, amelyeknek nincs saját "jelük". Más jelnyelvekben az ujjbetűzést csak ritkán használják, és az emberek inkább a teljes kifejezések és mondatok használatát részesítik előnyben. Számos kisebb jelnyelv létezik világszerte, amelyek különleges közösségekben alakultak ki és nincsenek befolyásolva a nagyobb jelnyelvektől. A Kolok Kata, más néven balinéz jelnyelv, egy falusi jelnyelv, amely két szomszédos faluban őshonos Bali északi részén, Indonéziában. A Kata Kolokban nincs hivatalos ujjbetűzés, ami rámutat arra, hogy az ujjbetűzés nem szükséges egy jelnyelv működéséhez. Ezek a kisebb jelnyelvek gyakran egyedülállóak és tükrözik az adott közösség kultúráját és történelmét.

Mint minden nyelvnek, a jelnyelveknek is van saját nyelvtana és szókincse. A szavak és mondatok jelentését a kézformák, helyzetek, mozgások és az arc kifejezései együttesen hordozzák. A nyelvtan komplex és nem feltétlenül követi a hallók által beszélt nyelv szerkezetét.

## Amerikai jelnyelv (ASL)

Az amerikai jelnyelv (ASL) az Egyesült Államokban és Kanadában élő hallássérült közösségek által használt nyelv. Az ASL története a 19. századig nyúlik vissza, amikor Thomas Gallaudet és Laurent Clerc megalapították az Egyesült Államok első hallássérült iskoláját. Az itt használt jelnyelv kombinálta az amerikai helyi jelnyelveket a francia jelnyelvvel, ami az ASL alapját képezte.

Közel félmillió ember beszéli, ami meglepő, hiszen ez a szám körülbelül 10%-a a 40 millió hallássérültnek az országban. Az ASL a 7. legtöbbet használt, viszont az interneten szabadon hozzáférhető források által legjobban dokumentált jelnyelv a világon, így a kutatás is ezzel foglalkozik. Fontos kiemelni azonban, hogy hasonlóan rögzített adathalmazok esetén a modellek, és algoritmusok univerzálisak.

### Ujjbetűzés az amerikai jelnyelvben

Mint sok más jelnyelvben, az amerikaiban is található ujjbetűzés. A brit jelnyelvvel ellentétben itt csupán egy kézre van szükség, a számok, illetve angol abc betűinek kommunikációjára. Ez persze nem azt jelenti, hogy az egyéb gesztusoknak nincs jelentése ebben a kontextusban, viszont a karakterek teljes mértékben beazonosíthatóak csupán a kezek megfigyelésével. pl.: A fejmozgás, vagy előre dőlő testhelyzet ujjbetűzés során hangsúlyozó szereppel bírhat.

Az ujjbetűzés olyan kézformákat használ, amelyek egyes betűket képviselnek, hogy kifejezzenek szavakat. Bár az ujjbetűzés csak az ASL egy része, használják nevek, címek, telefonszámok és egyéb, mobiltelefonon gyakran beírt információk közlésére. Sok siket okostelefon-felhasználó gyorsabban tudja ujjbetűzéssel kifejezni a szavakat mintsem, hogy begépelje őket. Valójában az ASL ujjbetűzése jelentősen gyorsabb lehet, mint a tipikus okostelefonos virtuális billentyűzeten való gépelés (átlagosan 57 szó/perc az amerikai 36 szó/perc átlagához képest).

A legtöbb gesztus statikus, vagyis nem szükséges dinamikus mozdulat a megkülönböztetéshez. Ez alól kivétel a j és z betűk, melyeket ugyan azzal a kéz pózzal kell jelelni, mint az i és d betűt, csupán a kézzel leírt alakzatban térnek el. Ezen felül a számok esetében előfordul, hogy a kéztartás megegyezik egy betűnél használttal. Megkülönböztetni őket a kontextusból lehet.

### Általános ASL

A legfontosabb különbség az ujjbetűzéshez képest, hogy sokkal fontosabb szerepet játszik a fej, és mimika, valamint egyéb gesztusok is. Egy jó példa erre a jelenségre az Anya, illetve Apa gesztus, melyek ugyanúgy széttárt tenyérrel viszont a fej különböző részét érintve fejezhetőek ki. Ez nem is egyedi eset, léteznek adatbázisok, amelyekben kéz alak alapján lehet gesztusokat keresni. Módosító tényező lehet még a másik kéz alakja, és a végrehajtott mozgáskombináció is.

A hagyományos angolhoz képest általában más a szórend. Tipikusan előre helyezi az idő vagy igeidő információt, majd a mondat témáját, és végül a megjegyzést. Például az "I am going to the store" (Én megyek a boltba) mondatot ASL-ben így jeleznénk: "NOW I STORE GO" (MOST ÉN BOLT MENNI), de az is gyakori, hogy az igét megismétlik a jelelés elején és végén. Ezen felül, mikor írásos formában akarjuk ábrázolni a jelnyelvet, még szokás vizuális annotációkkal is ellátni. Ezt nevezi az ASL „gloss” vagy „glossing” -nak. Ebben az olyan jelek is feltüntetésre kerülnek, melyeknek nincs angol megfelelője, illetve egyéb „non-verbális” gesztusok.

# Kapcsolódó kutatások

## Hagyományos képfeldolgozáson alapuló módszerek

Az ilyen módszeren alapuló megoldásokra igaz, hogy bár a sikeres projektek átalakítják a kamera képeit valamilyen magasabb szintű reprezentációba, de sosem lépnek ki a kép dimenzióból, nem modellezik a kéz fiziológiáját. Gyakori például a kezek szegmentálása. Vagy a dinamikus információ kinyerése az egyes képkockák pixeleinek különbségéből.

### Jelnyelv automatikus fordítása többfolyamos 3D CNN-nel és mesterséges mélységtérképek generálása

Ebben a kutatásban videó képkockáiból azonosítottak be jeleket. Első lépésként a képkockák számát csökkentették le. Egy előtanított VGG16 nevű modellt alkalmaztak a képkockák csoportosítására, majd K-means klaszterezés és Fő komponens analízis alkalmazásával 10 képet választanak ki a videóból. Erre a modellek és tanítási idő komplexitás csökkentése miatt van szükség. Ezt követően több érdekes algoritmust is kipróbáltak, egy újabb képi dimenzió bevezetésére a videó mellé.

* Optical flow, vagyis a tárgyak látszólagos mozgásának mintázatát közelítő módszert alkalmaztak az egyes képkockák között. Ehhez Gunnar Farneback 2003-ban publikált algoritmusát alkalmazták. Az így keletkező újabb képkockákat, hasonlóan a következő módszerhez, az RGB dimenzióhoz csatolták, egy újabb „optical flow” dimenzióként.
* A másik módszer mélység információt tartalmazó képek alkalmazása. Ehhez azonban nem egy második, mélységélességre is érzékeny kamerát alkalmaztak, hanem algoritmikusan generálták. A feladatot egy módosított GAN (Generative adversarial network) modellel végezték, melyben a szokásos zaj helyett, két bemenete van mind a generátornak, mind a diszkriminátornak, és melyen keresztül mindkét modell megkapja az eredeti képet is, így sikeres tanítás esetén lehetősége lesz a generátornak a bemenő képkockára illeszkedő mélység térképet generálni.

Az architektúrát tekintve a továbbiakban rétegekbe szervezett 3D konvolúciós blokkokat használtak, majd az eredményt 2 Dense réteggel osztályozzák.

A nehézséget, az ilyen jellegű megoldásoknál az okozza általában, hogy mivel képi adathalmazokról van szó, azok mérete és kezelése rendkívül költséges. Ez annak az eredménye, hogy a képek/videók sokkal több információt is hordoznak, mint amire szükség van a fordításhoz, és egyáltalán nem triviális olyan feldolgozó algoritmusokat találni, melyek optimálisak mind a tárol információ, mind az adatok tömörítése terén. A problémát fokozza, hogy a finom gesztusok, mint az arc mimikája, szemöldök mozgása, mivel minimális mozgással jár, megkívánja a nagy felbontású, jó minőségű képeket. Továbbá az is szembetűnő, hogy a kutatások jelentős része, kevés kategória közül képes gesztusokat azonosítani, ehhez persze a szűkös adathalmazok is hozzájárulnak, de az sem zárható ki, hogy a nagyobb kategória számhoz szükséges megnövekedett adathalmaz, és tanítási idők, korlátozzák a fejlesztőket.

## Segédeszközt használó megoldások

Segédeszköz használata a gesztusok felismerésére egy hatalmas előnnyel jár a hagyományos képfeldolgozást alkalmazó módszerekkel szemben, itt a gesztusok kategorizálásához különböző szenzorok alkalmazásával rendelkezésre áll valós térbeli információ a kezekről, valamint a mozgásokról. Ez hatalmas előnyt jelenthet a pontosság terén hiszen, kamerakép használatakor előfordulhat, hogy olyan szögben kerül rögzítésre a jelelő alany, amiből nem megállapítható a gesztus. Ez persze nem áll fent egy hétköznapi frontális kommunikáció, vagy az eszközeinkkel való interakció során, hisz ekkor egyéb iránt is a kamera felé orientálja magát a jelelő, de a segédeszközbe rejtett szenzorok minden esetben egy biztosabb, és pontosabb képed adnak a gesztusokról.

### Kesztyűt használó kutatások

Az egyik jelentős irány a kesztyű használata. Irodalma jelentős, néhány évente mindig érkezik egy-egy nagyobb áttörés a területen. A [korai prototípusok](https://ieeexplore.ieee.org/document/7785276) az egyszerűbb mozdulatok, például az ábécé betűinek és a számjegyeknek a felismerését célozzák általában, míg a [későbbi rendszerek](https://link.springer.com/article/10.1007/s40820-022-00887-5) bonyolultabb jelnyelvi struktúrákat és kifejezéseket is képesek kezelni.

A kesztyű valójában több szenzor összehangolt méréseit használja fel a gesztusok azonosítására. A hüvelykujj kivételével a maradék 4 ujj mozgatásának szabadságfokai limitáltak, elsődleges mozgási tartományuk a tenyér felé hajlítás. A dőlésszögek meghatározásának leggyakrabban használt érzékelője a hajlítás érzékelő szenzor, melynek ellenállása arányosan növekszik az ujjak mozgása során. Gyakran szerelnek továbbá gyorsulásmérőket az ujjak végére, valamint a tenyérre is. Nyomás érzékelőket az ujjbegyekre, valamint az egyik megoldás még a Hall effektust kihasználó távolság szenzor adatait is feldolgozza. Nyilvánvaló probléma viszont, hogy jeleléskor szükséges az eszközzel rendelkezni, ezáltal a természetességéből könnyen veszíthet a kommunikáció, valamint a minden kísérletben változó hardvereknek köszönhetően speciális adathalmazra van szükség, ami nem feltétlen robosztus.

### SignRing: Amerikai jelnyelv-felismerés IMU szenzorral ellátott gyűrűkkel

Egy újabb megközelítés a gyűrűbe helyezett inerciális mérőegység (IMU) használata. A hivatkozott kutatásban a két mutatóujjon elhelyezve 6 szabadsági fokos adatokból rekonstruálták a jelelt gesztusokat. Ez egy rendkívül friss, de ígéretesnek tűnő kutatás, mivel egy gyűrű viselete nem akadályozza az embereket a hétköznapokban, és javaslatot tesznek több szögből felvett videó esetén tanító adatok kinyerésére, mellyel a már meglévő videó alapú adathalmazokat is hasznosítani lehetne, viszont az általános felhasználhatósága egyelőre nem bizonyított. Ennek oka, hogy mint sok más jelnyelvhez hasonlóan, az amerikai változatban is, nem csak a kezek hordoznak információt. pl.: Azok helyzete az archoz képest jelentősen módosíthatja a gesztus jelentését. Mivel az arc szenzorokkal való monitorozása, végképp egy természet ellenes lehetőség, esetleg előzetes kalibrációra, vagy kontextusfüggő javításra van szükség a fordításhoz. Ezen felül az arc mimikája nélkül elveszítjük a jövőbeli lehetőséget a „gloss” teljes értékű fordítására. Így az ilyen módszereket nem tartom alkalmasnak univerzális megoldásként a jelnyelv fordítására.

## A dolgozat célja

A szövegbevitelhez szükséges jelnyelv-felismerő mesterséges intelligencia, bár a hivatkozott kutatásokból látszik, hogy sok fejlődést, és nagy potenciált rejt magában, még messze elmarad a hangról szöveggé átalakítás mögött, vagy akár a gesztus alapú gépelés mögött is. A dolgozat ehhez az aktívan kutatott területhez járul hozzá, mind az ujjbetűzés, mind az általános jelnyelv fordítása terén. További célkitűzés, hogy a fejezetben feltárt nehézségekre megoldást adjon:

* Mindennemű segédeszköz használata nélkül, a kéz fiziológiai modelljével dolgozzon.
* Univerzális, könnyen kezelhető adatokhalmazokat lehessen létrehozni a tanításra.
* A már meglévő videó/képi formátumú adatbázisok felhasználhatóak.
* Mindezt valós időkorlátok között.

# Megközelítés

Az előző fejezetben kiemelt nehézségeket egy újszerű képfeldolgozási módszerrel oldom meg. Ennek lényege, hogy 3 csatornás képekből előállítható, a képen látható kéz ízületeinek 3 dimenziós modellje. A módszer alkalmazható a test többi részén elhelyezkedő jellegzetes pontok beazonosítására, beleértve a fejet, és az arcot. A megfelelő pontok koordinátáit felhasználva kialakított modell alkalmas lehet a kitűzött összes cél megvalósítására. Konvertálhatóak a meglévő képi adatbázisok, a modellek futtatására felhasznált információ univerzális, ha előre megállapodunk mely jellegzetes pontokat használjuk fel, valamint rendkívüli mértékben tömöríthetné a képi adatokat.

A dolgozatban pózbecslés alapon működő modelleken kísérletezve vizsgálom, hogy az adatok ilyen formában való kódolása alkalmas-e a jelnyelv általános fordítására, illetve, hogy a jelenleg elérhető póz felismerő algoritmusok milyen pontosságú megoldásokat eredményeznek.

## Póz felismerés

A póz detektálás a számítógépes látás egyik kulcsfontosságú területe, amely az emberi test alakjának és tartásának felismerésére és nyomon követésére összpontosít. Ezen technológia segítségével a gépek képesek azonosítani és elemezni az emberi test különböző részeit, mint például a kezek, lábak, fej és a test főbb ízületei.

A működése lényegében képfeldolgozási és mély tanulási technikákra épül. A mély neurális hálózatok, különösen a konvolúciós neurális hálózatok (CNN), forradalmasították ezt a területet, lehetővé téve a nagy pontosságú és valós idejű póz detektálást. A hálózatok képesek "tanulni" az emberi anatómia jellemzőiről és azok változatosságáról nagy adathalmazokon keresztül, így képesek azonosítani a testrészeket és azok relatív elhelyezkedését.

Az elmúlt években jelentős előre lepések történtek ezen a területen. A modellarchitektúrák és az adatmegoldási technikák fejlődésével a póz detektálási rendszerek gyorsabbak és pontosabbak lettek. Ezen kívül az átfogó adathalmazok elérhetősége és az erősebb számítási kapacitások is hozzájárultak a terület fejlődéséhez.

### Open pose

Az OpenPose kiemelkedik a póz detektálás világában, mint az első nyílt forráskódú, valós idejű 2D test-, kéz- és arc-póz detektálási rendszer. A Carnegie Mellon Egyetem által kifejlesztett projekt olyan technológiát hozott el a közönségnek, amely a korábbi megoldásoknál jobban skálázható és sokoldalúbb. Míg számos rendszer létezik a póz detektálásra, az OpenPose különösen azzal emelkedik ki, hogy egyetlen neurális hálózattal képes kezelni a test, kéz és arc pózait, így integráltabb megoldást kínál. Továbbá, a nyílt forráskódú természetének köszönhetően sok kutató és fejlesztő számára elérhetővé vált, ami gyors innovációt és széleskörű alkalmazást tett lehetővé. Ezen funkciók kombinációja teszi az OpenPoset az iparág egyik vezető eszközévé a póz detektálásban.

A bemeneti kép először egy előtanított képfeldolgozó modellen megy keresztül. A következő lépésben egy kétfázisú konvolúciós blokkokból álló modell alkalmazása. Az első fázis végén hibafüggvénnyel kényszerítik a modellt, hogy egy úgynevezett végtag „affinitás” mezőt generálon. Ez a mező minden képponthoz egy 2 dimenziós vektort tartalmaz, ha adott pont rajta van a végtagon, akkor értéke a 2 legközelebbi felismerni kívánt kulcspont között feszülő vektor normálva, egyébiránt pedig 0. Természetesen a modell ezt csak megközelíti, de tanítás során egy olyan mezőt használnak „elvárt” adatként, ami az összes emberre kiszámolt mezők átlaga. A pontossága ellenőrizhető, két kulcspont között, egy integrállal a két pontot összekötő szakasz mentén.

A képen szöveg, Betűtípus, képernyőkép, sor látható

Automatikusan generált leírás

A képletben a kérdéses két kulcspontot jelöli, pedig tartományon a közöttük való lineáris interpoláció. megadja a keresett pontokon az affinitás mező értékeit, majd ez skalárisan szorzódik a két pont között feszülő vektor normáljával. Utóbbi művelet jól megfogja a kapcsolatot a két vektor hasonlósága között, hiszen értéke tartományon annál közelebb van 1-hez minél hasonlóbb irányba néznek.

A második iterációban az eredeti bemenet és a kiszámolt mező együttese halad tovább, egy hasonló konvolúciós architektúrába, ami pedig az egyes kulcspontok előfordulási valószínűségét adják. Több magas valószínűséggel rendelkező pozíció is lehetséges, hiszen a képen lévő emberek száma ismeretlen. Ekkor egy feladat maradt hátra, mégpedig az összetartozó kulcspontok azonosítása. A feladatot lépésekben oldja meg, egy gráfelméleti algoritmussal. Két kulcspont, ami között akarunk összeköttetést, vegyük például a vállat és a könyököt, képzeljünk el úgy, mint egy párosgráf csúcsai. (Egymással nem akarjuk különböző emberek vállát összekötni, így megvan a két csúcshalmaz.) Két csúcs közöti él súlya legyen a fenti integrállal kiszámolt érték. Ekkor a feladat enyhíthető egy maximális súlyú párosítás keresésével, amire Kőnig Dénes matematikus munkássága nyomán ismerünk polinomiális futásidejű algoritmust. Általános esetben szétválogatni a kulcspontokat változó számú emberre NP nehéz feladat lenne.

### Mediapipe holistic és hands

A MediaPipe Holistic azon kevés modell közé tartozik, amely egyesíti a különböző emberi testrészek detektálását, és még tovább megy: képes az egymáshoz képesti térbeli elhelyezkedésük azonosítására is. A Google által kifejlesztett MediaPipe platformon belül a Holistic modell olyan innovatív technológiákat alkalmaz, amelyek lehetővé teszik az arc, kéz és test pózának egyidejű elemzését.

Ami igazán különlegessé teszi a Holistic modellt, az az integrált megközelítés: több modellt kombinál együtt a lehető legpontosabb eredmények eléréséhez. Például külön modell végzi a testpóz, kéz, fej kulcspontjainak meghatározását. Ezek az almodelljei külön-külön is kiemelkedő teljesítményt nyújtanak területükön, de együtt alkalmazva a Holistic modell egy integrált és átfogó képet ad a felhasználó pózáról és interakciójáról környezetével.

Sajnos nem elérhető olyan részletes leírás az algoritmikus hátteréről, mint az OpenPosenak, de teljesítménye magáért beszél. Csupán egy embert képes felismerni egy képkockán, de ez egyáltalán nem korlátozó jelenleg, hiszen a feladat egy időpillanatban csak egy jelelő fordítására terjed ki. Előny viszont, hogy nem csak 2D koordinátákat ad vissza, hanem térbeli fogalommal is rendelkezik az ember pozíciójáról. Ez a kezek, jelnyelv céljával való feldolgozása szempontjából hatalmas előnyt jelent. További pozitív tulajdonság, hogy az könyvtár, és vele együtt a modellek könnyen futtathatóak a különböző platformokon. Elérhető például Androidon, illetve böngészős környezetben is. A valós idejű futás hasonló erőforrás korlátozott környezetekben algoritmikusan támogatott. Ha folyamatos videóból hajt végre póz felismerést, abban az esetben az egyes képkockák között kihasználja a lokalitásokat, vagyis azt a tényt, hogy a végtagok nem tudnak hirtelen túl nagy távot megtenni fizikai korlátaik miatt, így egy kisebb komplexitású követő algoritmust alkalmaz.

Az imént felsorolt előnyök miatt a dolgozat további részében a mediapipe holistic, illetve a kézre dedikált verzióját (mediapipe hands) fogom használni.

# Statikus ujjbetűzés

Két alapvető módszer különíthető el a jelnyelv-felismerés terén: a dinamikus videóból és a statikus képkockákból történő fordítás. Míg a dinamikus videóból történő fordítás lehetővé teszi az egymást követő mozdulatok és azok közti átmenetek azonosítását, ami gyakran kulcsfontosságú a jelentés teljes megértéséhez, a statikus képkockákon alapuló fordítás kihívások elé néz. Ebben az esetben a rendszernek egyetlen pillanatfelvétel alapján kell meghatároznia a mozdulat jelentését, ami korlátozott információt kínál a mozdulat teljes kontextusáról. Ennek ellenére mindkét módszernek megvannak a maga előnyei és alkalmazási területei, amelyekről mélyebben is érdemes beszélni.

A statikus módszer nem véletlen az ujjbetűzés témakörében jön elő, ugyanis a j és z kivételével egyik gesztus sem kíván dinamikus mozdulatot, ez a kettő pedig egy későbbi szakaszban tárgyalt algoritmussal korrigálható. Egy másik probléma is fennáll, ha az 1-9-ig lévő számjegyeket is bevonjuk. Az ilyen „kiterjesztett” ujjbetűzés ugyanis, tartalmaz olyan kéztartásokat, melyek két kategóriába is tartoznak. Ilyen például a 2 és v, melyek gyakorlatilag identikusok kéztartás szempontjából. A gyakorlatban szövegkörnyezetből megkülönböztethetőek, de ha mégis szükség van izolált környezetben a használatukra, olyankor szám esetében a jelelő tenyerét maga felé fordíthatja.

## Adathalmaz

Mivel a kutatás időpontjában nem állt rendelkezésre statikus póz adathalmaz a feladatra, így először képi adatbázisokat kerestem, hogy az adatokat konvertálás során megfelelő formátumúra alakítsam. Mivel az erre a feladatra kiélezett adathalmazok csupán kivágott kezeket tartalmaznak, és a mediapipe holistic modell csak az emberi test kontextusában képes részleteket felismerni, így kézen fekvő volt a mediapipe hands api használata. A fellelhető adathalmazokban közös volt, hogy a képeket erősen előfeldolgozták, kis felbontásúvá konvertálták és a legtöbb esetben még a szín információt is elhagyták. Pontos specifikáció nem érhető el az api minimális minőségi követelményeiről, így több, egymástól merőben eltérő, nem csak ASL adathalmazt is kipróbáltam:

1. Sign language MNIST: A klasszikus kézzel írt, 28x28 pixeles számjegyeket tartalmazó MNIST adathalmaz variánsa, amerikai jelnyelv feladatra specializálva. A képek hasonlóan névrokonához fekete fehérek, és identikus felbontásúak.
2. University of Exter ASL dataset: Két féle verzióban is elérhető. Tartalmaz alacsony felbontású színes képeket, valamint mélység információt is, bár utóbbi nem került feldolgozásra. A képek minősége elég változatos, és majdnem mindegyik tartalmaz artifaktumokat.
3. Sign language for Alphabets adathalmaz: Szintén fekete fehér adathalmaz, de magasabb felbontású, mint az első pontban, viszont nem amerikai jelnyelv.

Első körben az MNIST adathalmazzal kezdtem dolgozni. Nyers formában a képek nem voltak megfelelő minőségűek, a mediapipe hands api számára, így különböző minőség fokozó technikákat alkalmaztam rajtuk.

### Interpolációs algoritmusok

Az elsődleges gyanúm a képek kis felbontása volt, ezért először ezt próbáltam orvosolni. Az interpolációs algoritmusok kulcsszerepet játszanak a képek felbontásának javításában. Az interpoláció lényegében magasabb felbontású területeken pixelértékek becslését jelenti az alacsony felbontású képről származó információk alapján. A klasszikus interpolációs módszerek, matematikai összefüggések használatával számolják ki ezeket az értékeket, figyelembe véve minden pixel közvetlen szomszédjait. Azonban ezek a hagyományos módszerek néha nem tudják megragadni a bonyolult részleteket. A mélytanulás megjelenésével fejlettebb interpolációs algoritmusokat fejlesztettek ki, amelyek neurális hálózatokat használnak. Ezek a modellek, alacsony és magas felbontású képek párosairól tanulnak, és céljuk, hogy hidat képezzenek az egyszerű matematikai becslések és a valós világ képeinek összetettsége között, így kínálva pontosabb és esztétikailag kellemesebb szuperfelbontási eredményeket.

#### Hagyományos matematikai módszerek

Az interpoláció során a matematikai módszerek a pixelértékek közötti összefüggéseket használják fel a becslésekhez, és nem igényelnek külső információs forrást. A kipróbált módszerek:

1. Legközelebbi szomszéd (Nearest Neighbor): Az egyik legegyszerűbb módszer. Itt az új pixel értékét közvetlenül az eredeti képen található legközelebbi pixel értékéből veszi. Nincs súlyozás vagy más számítás, így gyorsan működik, de a végeredmény gyakran "kockás" megjelenésű lehet, különösen nagyobb méretarányú nagyításnál.
2. Lineáris (Bilinear) Interpoláció: Négy közeli pixel értékét veszi figyelembe. A kimeneti pixel értéke a környező pixelértékek súlyozott átlagaként jön létre, ahol a súlyok az új pixel távolságán alapulnak a négy eredeti pixelhez képest. A bilineáris interpoláció simább képet eredményez, mint a legközelebbi szomszéd módszere.
3. Bikubikus Interpoláció: Két érték interpolációját egy harmadfokú polinóm illesztésével valósítja meg. Hogy a paraméterei egyértelműen kiszámíthatóak legyenek szükséges a végpontokban vett deriváltak értéke, ennek ismeretéhez pedig a környező szomszédok is. Így egy pixel értékének kiszámításakor figyelembe veszi a környező 16 pixel értékeit. Eredménye sokkal természetesebb kinézetet ad, mint a lineáris változat.
4. Lánczos Interpoláció: A módszer egy magyar matematikus, Lánczos Kornél érdeme. Egy ablakozott sinc-függvényt használ a pixelek súlyainak kiszámítására, és több pixel értékét veszi figyelembe, mint a bilineáris vagy bikubikus módszerek. A használt (opencv) library a pixel 8x8-as környezetét veszi figyelembe. A sinc-függvény tulajdonságainak köszönhetően a Lanczos kiválóan alkalmas képek élesítésére anélkül, hogy jelentős torzulást okozna.

#### Mély tanulás alapú módszerek

A mélytanulás-alapú interpolációs technikák számos előnnyel rendelkeznek a hagyományos matematikai módszerekhez képest. Elsődlegesen ezek az algoritmusok képesek észlelni és reprodukálni azokat a bonyolult mintákat és szerkezeteket a képeken, amelyeket a klasszikus módszerek esetleg nem vesznek észre. A hagyományos módszerek rögzített matematikai képletekhez kötöttek, amelyek néha nem képesek alkalmazkodni a valóságos képtartalom széles változatosságához. Ezzel szemben a mély tanulás alapú módszerek hatalmas adatkészleteken tanulnak, lehetővé téve számukra, hogy általánosítsanak és tájékozottabb előrejelzéseket tegyenek a számtalan képminta alapján, amelyekkel találkoztak. Továbbá, az ilyen modellek képesek hierarchikus jellemzőket felfogni, ami azt jelenti, hogy észlelni tudják a képek alacsony szintű részleteit, mint a szélek és textúrák, valamint a magas szintű tulajdonságokat, mint az objektum struktúrák és szemantikai kapcsolatok. Ez a holisztikus megértés gyakran olyan interpolált képekhez vezet, amelyek nem csak élesebbek, hanem tartalmukban is koherensebbek, csökkentve az artifaktumokat, olyan vizuális eredményeket előállítva, amelyek közelebb állnak ahhoz, amit az emberi szem elvárna látni. A kísérletek két modellel is el lettek végezve:

1. A „Densely Residual Laplacian Super-Resolution” tanulmány alapján készített modell, továbbiakban DRLN, különlegességét az adja, hogy olyan modulokból épül fel, melyek kombinálják a mély reziduális kapcsolatokat, vagyis modulon belül a konvolúciós blokkok minden őket megelőző blokk kimenetét megkapják bemenetükön, valamint egy speciális Laplace figyelmi mechanizmust. Továbbá ezek a modulok is többször szerepelnek egymás után. Paraméterek számában mérve komplexebb, valamint futásidőben jelentősen lassabb, mint a következő tesztalany.
2. Az „Attention in Attention Network for Image Super-Resolution” című írás alapján készült modell, röviden A2N, erejét szintén meglévő elemek újszerű felhasználásból nyeri. Alapja egy úgynevezett „Attention in attention block”, mely lényegében dinamikusan tanult súlyokkal kombinálja a bemenet általánosan feldolgozott, valamint figyelmi mechanizmussal ellátott ágát.

### Fekete fehér képek színezése mély tanuláson alapuló modellekkel

A képek felbontásának növelése nem járt átütő sikerrel a vizuális felismerhetőség ellenére, így csak egy tulajdonság hiánya okozhatja a mediapipe könyvtár felsimerési próblémáját, mégpedig a képek színe.

Szerencsére túl vagyunk már azon az időszakon, mikor fekete fehér képeket manuálisan kellett megszínezni. Léteznek ugyanis előtanított mesterséges intelligencia modellek, melyek fekete fehér képekből színeket javasolni, minden pixelnek. Ez úgy lehetséges, hogy sok tanítókép alapján a modellek megtanulják értelmezni a képeken szereplő objektumokat, és hogy a nagy számok törvénye alapján hogyan néznek ki színesben.

A felhasznált modellek a „Colorful Image Colorization” című tanulmány alapján készültek. Alap gondolat, hogy a képeket váltsuk át „LAB” formátumúra. A modellek bemenetként a fénysűrűség (Luminance) csatorna értékeit kapják, és az AB-csatorna értékeit próbálják közelíteni. Egy fekete fehér képen, a tárgyak nagy része több színt is felvehet így, ha regressziót alkalmazunk például euklideszi távolságmetrikával, az átlagoló hatás fakó színek felé tolja el a matematikailag optimális eredményeket. A szerzők javaslata, hogy a 10 egységenként felosztott AB-csatorna diszkrét egységeire alkalmazható klasszifikáció, élénkebb színeket eredményez. Adathalmazként felhasználható bármilyen színes kép, színtér váltás után rendelkezésre is áll a bemenet, és elvárt kimenet. A továbbiakban ECCV16 néven hivatkozott modell ezen az elven lett betanítva.

Kipróbáltam egy alternatív módon betanított modellt is (SIGGRAPH17). Az architektúra ugyanaz maradt, de alkalmanként tanulás során adunk a modellnek néhány képpontban segítséget az elvárt színhez. Bár a kutatásaim során ezt nem használtam ki, de így utólag is lehetőség nyílik segíteni a modell színválasztását. A tanítás során megadott segédpontok számát egy geometriai eloszlás függvény határozza meg, így bőven akad olyan bemenet is, amin nincsen „segítség”, így alkalmas a felhasznált feladatra. Tapasztalat szerzés céljából, néhány példán futtatva, bár kevésbé élénk színeket produkál, mégis realisztikusabb képeket eredményez, mint a társa.

### Minőség javító algoritmusok összehasonlítása

Bár a felbontás növelő technikák látványos eredményt értek el, amíg a képek fekete fehérek maradtak nem volt képes a mediapipe api kezeket felismerni. A színezési technikák közül az ECCV16 mutatta a legjobb eredményeket. Ez a modell nem mellesleg a használt neurális háló kialakítása miatt változtatja a bemeneti képet méreteit. Előfeldolgozásként a Lánczos interpolációval megnöveli a kép méreteit 256x256 pixelre, majd az inferencia után Lineáris interpolációt alkalmaz az eredeti méretre való összenyomáshoz. Végeztem teszteket a felbontás növelés és színezés egymás utáni alkalmazására, ezekben az esetekben eredményeket szintén azok a kombinációk hoztak, melyekben szerepelt az ECCV16 algoritmus, de a felismert póz közel minden esetben erősen torzított lett. Az eredményeket validáltam a korábban „Sign language for Alphabets” néven említett adathalmazon is. A teljes ASL-MNIST adatbázist konvertáltam az imént említett színező modellel, ezzel az esetek 30%-ában, közel 8500 esetben, sikerült pózt felismerni a képeken.

A minőség javító algoritmusokat a „University of Exter ASL” névvel illetett adathalmazon is kipróbáltam. Mivel ez egy színes képeket tartalmazó adathalmaz, a színező algoritmusok nem segítették a találati arányt, de a felbontás növelése Lineáris interpolációval igen. Konvertálás után az eredeti adathalmaz 38%-át sikerült megőrizni.

### Saját ASL adathalmaz

A más adathalmazokon elért sikerek ellenére, saját ASL ujjbetűzés adathalmazt gyűjtöttem. Az adatok frontális irányba néző webkamerával kerültek rögzítésre. Az adatokat 2 különböző napszakban, 2 különböző helyszínen készültek. A felhasznált minták között egyénenként egyenletesen oszlott meg a különböző kategóriákban rögzített felvételek száma, valamint a jobb és bal kéz használata.

A nyers felvételek nem, csak a póz információ került mentésre. A képeken maximum egy kezet kerestem, mivel az ASL-ben minden betű és szám kifejezhető így. Képkockánként 21 db 3 dimenziós kulcspont került rögzítésre. Feltételezzünk 32 bites lebegőpontos számábrázolást, így fevételenként 252 byte adattal kell számolni. Ha el akarnánk tárolni a képeket, erős előfeldolgozás mellett is, például 100x100-as méretű, 8 bites színmélységű, fekete fehér képeket esetén 10000 byteot kellene tárolni, ami így közel 1: 40 tömörítési arányt jelent.

Mivel az adatok átlagosan 15-20 fps sebsséggel kerültek rögzítésre, még a kezek gyors mozgásának hatására is sok hasonló kép keletkezett, ezért 10 egymást követő adatpontból 9 nem került felhasználásra, így biztosítva, hogy legalább 0,5 sec teljen el két minta között.

A külső forrásból származó képekkel ellentétben nem vágtam ki, és izoláltam a kezeket a póz információ kinyerése előtt. Ez két okból is fontos lehet. Egyrészt az póz felismerő api relatív koordinátákat ad vissza. Ezek az értékek 0 és 1 közé esnek, és azt jelzik, hogy a bemeneti képen a szélesség, és magasság hány százalékánál találhatók az adott pontok. Ha a bemeneti kép más képarányú, a detektált kezek koordinátái elnyúltnak torzítottnak tűnhetnek egymáshoz képest. Hasonló jelenség figyelhető meg a különböző látószögű kamerákon keresztül felismert kezeken is. A másik jelentős különbség, hogy mivel a teljes kép csak egy kis részén helyezkednek el a kéz a koordinátái matematikailag sokkal kisebb mértékben térnek el egymástól, mint a korábban tárgyalt adathalmazok esetében.

## Modellek

A fejlesztés során egy nyílt forráskódú könyvtárat, a TensorFlow-t használtam, amelyet mély tanulási modellek fejlesztésére és betanítására terveztek. A TensorFlow-ban található Keras könyvtár magas szintű API-kat kínál, amelyek közül az egyik a Sequential API. Ez az egyik leginkább egyszerűsített és struktúrált módja a neurális hálózatok létrehozásának a TensorFlow-ban. Ez az API lehetővé teszi a rétegek egymás utáni hozzáadását egy lineáris stackben, ami ideális az egyszerű előrecsatolt neurális hálózatok modellezéséhez.

A klasszifikációs feladatként megfogalmazott problémára két féle modellt próbáltam ki:

1. Egyszerű mély neurális háló.
2. Konvolúciós rétegekkel kiegészített CNN architektúra.

Mivel nem képi adatokról van szó, nem feltétlenül indokolt szemantikusan a konvolúciós rétegek használata. Az építőelem eredeti célja, hogy lehetővé tegye a modellek számára térbeli hierarchia megtanulását, valamint, hogy az ilyen jellegű mintákat pozíció invariánsan fel tudják ismerni. Belátható, hogy a póz információra ezek az eshetőségek ugyanúgy fennálnak, és az eredmények igazolják is az intuíciót. További kiegészítésként „Dropout” rétegeket is helyeztem el a hálózatokban, melyek bizonyos valószínűséggel néhány réteg bemenetét 0-ra állítják. Ez egy gyakran alkalmazott technika a modellek általánosító képességének fejlesztésére.

Az így megfogyatkozott adatpontokat minden epoch során egy bizonyos valószínűséggel affin transzformálom. Ez segít a modell általánosító képességének növelésében. A transzformáció csak az xy koordinátákra érvényesül, ugyanis tapasztalati alapon kiderült, hogy a z irányú adatok nagysága nincs korrelációban a kéz fiziológiájával, nem arányos a másik 2 tengely menti kiterjedésével. Csupán annak eldöntésére alkalmas, hogy az egyes ujjak hogyan helyezkednek el térben egymáshoz képest. A bal kézzel jelelők, a jobb kézzel jelelők tükörképei, így bár bal kezes adatok is rendelkezésre állnak, mégis alkalmanként tükrözés történik az y tengelyre. További transzformációk között szerepel a képernyő síkjában történő közelítés, illetve távolítás, xy tengelyek menti nyírás, valamint 10 fokkal való elforgatás. A transzformációk paramétereit empirikus alapon határoztam meg, nagy számú mintán való teszteléssel, a túlzott torzítás elkerülés érdekében.

Eredmény grafikonok, teszt adathalmazon elért eredmény, conf mat

## Kiegészítő algoritmusok

A mesterséges intelligencia (AI) modellek fejlesztésekor a legoptimálisabb teljesítmény és pontosság elérése érdekében gyakran szükség van kiegészítő algoritmusokra. Blackbox jellegükből fakadóan, valószínűségi alapon viselkednek éles adatokon. Előfordulhat például, hogy bizonyos szögben, vagy a környezeti viszonyok hatására, pl.: fény csillanás a kamerában, nem lesz megfelelő a pillanatnyi kimenet. Ez a jelenség egyszerűen kiszűrhető, a többségi döntés algoritmusával.

### Kimenet korrigálása valószínűségi alapon

A modell eredményei először egy sorba kerülnek, ahonnan többségi döntés alapján kerül ki a modell valós predikciója. Egyszerűen megvalósítható egy fifó adatszerkezettel, és így hatékony módon kiszűrhetőek a pillanatnyi anomáliák. Szűrhető továbbá az a jelenség is, amikor ismeretlen pózok eredményéül kis magabiztosságú, csapongó eredményeket ad a modell, ilyenkor nem egyszerű többségi döntést hozunk, hanem csak egy bizonyos százalékú többség hatására adunk eredményt pl.: ha 70% „fölénnyel” egyik kategória sem rendelkezik, abban az esetben sikeresen azonosítottuk azt a helyzetet, amikor a modell képtelen a magabiztos azonosításra.

Ennek a megoldásnak egy hátránya van, mégpedig az, hogy nem veszi figyelembe a modell magabiztosságát, a tanítás során ugyanis valószínűségi értékeket is megtanul. Ha például a kimenetéből az látszik, hogy a bemeneten lévő póz 3 karakterhez tartozhat 90%-ban, és mindegyiknek nagyságrendileg egyforma az esélye, pontatlanság az egyiket győztesként kikiáltani és behelyezni a sorba. Egy ilyen szituációban, ha lemerevítené a kezét a felhasználó, a modell ugyanazt a feltételezhetően rossz megoldást adná vissza, és a többségi döntést kijátszhatná. Egy gyakran alkalmazott technika ilyen esetben, hogy vesszük a modell javaslatait, és valószínűségeik szerint csökkenő sorrendbe rendezzük őket, majd annyit választunk a legnagyobbakból sorban, hogy a valószínűségük összege ne legyen több egy p paraméternél, de maximum k darabot választhatunk. A választásokat átskálázzuk, hogy a kisebb csoportban is 1 legyen az összegük, és valószínűségi alapon választunk közülük elemet a bufferbe. Ha a modell elég magabiztos a választásaiban, a probabilisztikus viselkedésből fakadó „melléfogásokat” elfedi a felhasználó elől a többségi döntés, de kevésbé magabiztos esetben pedig valószínűségi alapon döntünk a javasolt kategóriák közül. Ezen technika a kategóriák számának növelésével egyre jobban működik, valamint a top p és k paraméterek a modell úgynevezett „kreativitását” befolyásolják. Előszeretettel alkalmazzák nagy nyelvi modellek esetében.

### Nem szándékos mozgások elkülönítése

Eddig feltételeztük, hogy a felhasználó minden időpillanatban jelel. Ez a valóságban nem áll fenn, és szükség van ennek megbízható detekciójára.

Első lehetőség egy okos heurisztika alkalmazása. Ha például nem érzékel kezeket a mediapipe api, biztosan nem jelel a felhasználó, továbbá egy lépéssel tovább is vihető ez az ötlet, azzal az információval kiegészítve, hogy a jelelés legtöbb esetben a test felső harmada körül, de a fejtől kissé lentebb történik. Ha viszont a rendszert például online konferenciák, megbeszélések spektrumában is el akarjuk képzelni, bizonyára elő fog fordulni, hogy az emberek akaratlanul is könyökölnek az asztalon, vagy az arcukhoz érnek. Az ilyen véletlen mozgások, valamint egyéb különleges testhelyzetek kiszűrésére nem elég tehát egy egyszerű pozíció heurisztika.

Lehetőség van egy újabb, „ismeretlen gesztus” kategória bevezetésére a klasszifikációs feladathoz. Ehhez a saját adathalmazomat kiegészítettem olyan felvételekkel, melyen jeleléshez hasonló, esetenként kényszeres cselekvéseket hajtok végre, mint például fej vakarás, könyöklés az asztalon. Az így újra tanított modell képes lesz elkülöníteni néhány esetben a nem jeleléshez tartozó mozgásokat, viszont problémája lesz az olyan esetekkel, mikor a mozgása során, bár nem jelel, mégis felvesz olyan pozíciókat a keze, melyek megfelelnek egy-egy kategóriának. Nem elég tehát a statikus pillanatfelvétel a feladat elvégzésére. A további fejezetek részletesen foglalkoznak olyan módszerek kialakításával, amelyek képesek kezelni ezt az elvi lehetőséget is, adok azonban egy ideiglenes megoldást is:

A javaslat egy másik kiegészítő modell alkalmazása, amely bináris klasszifikációt hajt végre minden időpillanatban, a jelelés tényének megállapítására. Bemenete az elmúlt n db időpillanatból származó információmorzsa is. Adatként ingyenesen elérhető videókat használtam fel, melyeket lerövidítettem, hogy csupán a tényleges jelelést tartalmazzák. A jelelést nem tartalmazó videó kialakítása kissé nehézkesebb, mert nagyon sok lehetséges elfoglaltság, és mozgáskombináció létezik. Választás során olyan szemlélettel kerestem a negatív kategóriába videókat, melyek ülő pozícióban, esetleges konferenciahívás alkalmával fordulhatnak elő. Ebben a feladatban már felhasználtam a testpóz approximációs modelleket is, ugyanis mint azt már korábban említettem a jelelés egy bizonyos testtartás pozícióban szokott történni, és ennek észlelésére hasznos lehet a kezeken kívül más adat is. Az előfeldolgozási lépések megegyeztek a klasszifikációs modell esetében alkalmazottakkal. Az n paramétert, mint a bemeneten szereplő elmúlt képkockák száma, empirikus módon hangolva 15 -nek állapítottam meg. A modell architektúrájához LSTM (Long Short term memory) elemeket használtam fel. A tanulmány nem foglalkozik, ezen modell optimalizációjával, de a szekvenciafeldolgozással kapcsolatos mély tanulási apparátus lehetőségeit egy későbbi fejezet külön tárgyalja.

Az eddig tárgyalt összetett modell betűzést képes felismerni. Másodpercenként 30 képkocka esetén 30 megfejtett betű keletkezik. Ha ezeket szavakká akarjuk összefűzni elég eltárolni az előző megfejtett betűt, majd csak a következő ettől különbözőt bevenni a megfejtések közé. Ujjbetűzést legtöbbször összefüggő kifejezésekre alkalmazzák, de előfordulhat, hogy külön akarjuk választani a szavakat. Ideiglenes megoldásként a szavak határait a jelelés pillanatnyi szüneteltetése jelenti.

### Hibák korrekciója nagy nyelvi modellekkel

Az eddigiekben nem esett szó a j és z betűkről. Ezek ugyanis mozgást igényelnek, és statikus képekből nem különíthetőek el az i és d betűktől. Ezeket a jeleket a modell tervezett módon el fogja rontani, így szükséges egy újabb kiegészítő algoritmus vagy modell, ami korrigálja őket. Első ötletként lehetséges megfigyelni az elmúlt néhány képkockán a kezek mozgását, és ez alapján megkülönböztetni az eseteket. Ez az amerikai jelnyelv esetében működhet, de koránt sem egyszerűen általánosítható más nyelvekre, ahol esetleg jelentősen több a mozgást igénylő jel. További gond, hogy a szavakat is betűző modell jelenleg nem tud dupla betűket adni a kimenetén.

Szöveges tartalmak javítására alkalmazhatóak a nagy nyelvi modellek képességei. Bemenetként egy empirikus módon kialakított utasítás (prompt) szolgál, mely a modellt arra utasítja, hogy javítsa ki a fent említett hibákat. Az irodalom ezt zero shot learningnek nevezi, ilyenkor gyakorlatilag egyetlen utasításból „tanul” a modell mindenféle példa ismerete nélkül. Egy fokkal kifinomultabb módszer a few shot learning, amely során példákat is elhelyezünk a prompt-ban. Ez gyakran segít a kimenet megfelelő strukturáltságának kialakítását is.

Ez az algoritmus rendkívül költség hatékony lehet egy külső nagy nyelvi modell szolgáltatás igénybevételekor, hiszen sok fejlesztési időt nem igényel, könnyen adaptálható egyéb hibák javítására is, nem beszélve más jelnyelvek integrációjáról. Hátrány viszont, hogy semmi féle minőségi garanciát nem vállal a kimenet helyességére.

## Értékelés

A kialakított modell, és algoritmikus apparátus alkalmas különálló szavak betűzésére, és korlátok között összetettebb kifejezésekre is. A különböző adathalmazok kombinációjaként előálló robosztus tanító adatmennyiség alkalmas a modell olyan szintű betanítására, hogy ismeretlen fényviszonyokban, változó kézméretű alanyok, a kamerától eltérő távolságból is kényelmesen használni tudják.

Izolált karakterek felismerésére tökéletes megoldás lenne, ha nem léteznének mozgást igénylő gesztusok, illetve a számokat bevonva identikus kéztartást használó jelek. Szavakká való összefűzés esetén ugyan bizonyos mértékig lehetőség nyílik ezen pontatlanságok korrigálására, de a jelek közötti váltásokból adódó hibák korrekciójára kialakított csúszó ablak méreténél fogva egy körülbelül 1 másodperces késleltetést rak a rendszerre karakterenként. Ez kezdők számára nem feltétlenül probléma, de haladó jelelők képesek ezzel összemérhető idő alatt jelelni, illetve egyesek akár gyorsabban is, ami kellemetlenségeket okozhat. Ez a probléma feloldható lenne, ha a modellek bemeneteként nem csak egy képkockáról származó adat kerülne, hanem egy bizonyos ablakméret keretein belül a múltbéli képekről is állna rendelkezésre információ. A további fejezetek ilyen modellek kialakításával foglalkoznak.

# Ujjbetűzés szekvenciális bemenetből

A jelnyelv-felismerés területén a statikus képkocka-alapú módszereken túl a szekvenciális adatokból történő felismerés is jelentős szerepet tölt be. Ez a módszer nem egyetlen pillanatfelvétel alapján próbálja megfejteni a gesztus jelentését, hanem a gesztusok egymás utáni sorozatát elemzi. A szekvenciális módszerek képesek felfogni a mozdulatok közötti átmeneteket, ami különösen fontos azoknál a jeleknél, amelyek hasonló alapmozdulatokkal kezdődnek, de eltérő befejező mozdulatokkal végződnek. Ez nem csak a pontosságot növeli, hanem a sebességet is, hiszen szekvenciális bemenet esetén elhagyható a kimenetet finomító csúszó ablak. Az egymás utáni képkockák elemzése lehetővé teszi, hogy a rendszer érzékelje a mozdulatok irányát és sebességét, ami a statikus képkockákban nem lenne látható. Ez lehetővé teszi a kifinomultabb jelentésátadást, ahol az átmenetek és a mozdulatok sebessége, ritmusa vagy dinamikája is hozzájárul a jelentéshez.

Összességében a szekvenciális adatokból dolgozó módszerek egy mélyebb és árnyaltabb megértést kínálnak a jelnyelv és az ujjbetűzés dinamikus aspektusairól, és azokról a finom részletekről, amelyek a kommunikáció során kulcsfontosságúak lehetnek. Ezen fejezet célja egy olyan módszer kialakítása, mely képes a dinamikát igénylő jelek (j, z) megkülönböztetésére, valamint az egyező kéztartást igénylő betű és szám párosok megkülönböztetésére kontextus alapján, mindezt növekedett sebességgel a statikus módszerekhez képest.

## Adathalmaz

Az adathalmaz egy Google által meghirdetett Kaggle versenyből származik. A Kaggle egy online platform, amely adattudományi versenyeket szervez, és lehetővé teszi a felhasználók számára, hogy kollaboratív módon osszák meg egymással adathalmazaikat. A verseny célja az Amerikai Jelnyelv (ASL) ujjbetűzésének észlelése és szöveggé történő fordítása volt. Az adatok több mint hárommillió ujjbetűzött karaktert tartalmaz, amelyeket több mint 100 siket jelnyelv-használó készített okostelefon kamerával, különféle háttér és világítási feltételek mellett.

A jelelőknek szöveget mutattak, amely nagybetűket is tartalmazott, azonban a legtöbben ezt nem közvetítették. Néhány minta azonban tartalmaz ezzel kapcsolatos technikákat:

* Görbe-L kézformát használva a nagybetűsítendő betű előtt.
* A nagybetűket fizikailag magasabban helyezve el a térben, mint a kisbetűket.
* Betű kézformájának megrázásával jelezve.

A készítő sajnos a címkéken erre utaló információt nem helyezett el, így a saját megoldásomban erre nem térek ki.

Két különböző verzió érhető el az adathalmazon belül. Az egyikben linkek, címek telefonszámok jelelt felvételei találhatóak. A másik összefüggő szöveges mondatokat tartalmaz. Ez az ujjbetűzés területén elérhető legnagyobb póz adathalmaz.

Preprocess

statistics

## Modellek

A feladatban folyamatos jelelést kell szöveggé alakítani. Ez egy szekvencia-szekvencia feladat, ahol mind a bemenet, mind a kimenet egy folyam. A feladat sokban hasonlít a hagyományos fordításhoz, azzal a kivétellel, hogy a kettő szekvencia reprezentációja, és felépítésre itt merőben különbözik. Ennek ellenére egy az általános nyelvfordítás területén gyakran alkalmazott encoder-decoder architektúrát választottam.

### Encoder-Decoder architektúra

Az encoder-decoder modell egy olyan neurális hálózati struktúra, amely a természetes nyelvfeldolgozás (NLP), az idősoros predikció, és más feladatok megoldására szolgál. Az ilyen modell két fő részből áll: az encoderből és a decoderből. Az encoder feladata, hogy magas dimenziós bemeneti adatokat alacsony dimenziós reprezentációvá alakítson. Ezt az alacsony dimenziós reprezentációt általában látens térnek nevezik, amely segít az információ tömörítésében és az esszenciális jellemzők megtartásában. Az encoder hálózat tehát z bemeneti adatokat elemzi, és egy kontextusvektort hoz létre, amely tartalmazza a bemeneti adatok alapvető információit. A decoder ezt a vektort veszi alapul, és kísérli meg rekonstruálni vagy előállítani a kívánt kimeneti adatokat. A decoder egy másik neurális hálózat, amely a kontextusvektort használja, hogy a kimeneti adatokat létrehozza vagy előrejelezze. Gyakran előfordul az is, főként fordítási feladatokban, hogy a decoder több részletben fejti meg a kimenetet, például betűnként, vagy szavanként. Ilyenkor a kontextus vektor mellett a bemenetén megjelenik az előző kimeneteinek összege. Erre is gyakran hivatkoznak kontextus néven, ezért a félreértések elkerülése végett jelen írásban az encoder kimeneteként előállt kontextus vektort encoder kontextusként, míg a decoder bemenetén megjelenő múltbéli kimenetét decoder kontextusként hivatkozom. Az egységet amire felosztjuk a decoder be és kimenetét token-nek nevezzük. Egy példával élve, angol magyar fordítás területén ahelyett, hogy a „science” szóból egy lépésben „tudomány” jelenjen meg a kimeneten, betűnként futtatjuk a decoder-t, mely megkapja a tudomány szó részleteit is bemenetként a kontextus vektor mellett, és minden alkalommal csak a következő betűről kell nyilatkoznia. Ez egyrészt azért előnyös, mert kevesebb az összes lehetséges token, ugyanis betűből kevesebb van, mint tetszőleges kombinációikból, így egyszerűbb a feladat, hiszen kevesebb kategória közül egyszerűbben tanul a modell választani. Másrészt a decoder kontextus sokat segíthet a kimenet formálásában, vegyük például az amerikai jelnyelvben a számok ujjbetűzését: nem minden esetben egyértelmű, hogy a jelelő a számra, vagy a betűre gondol, hiszen ugyanaz a kéztartás szükséges mindkettőhöz viszont, ha telefonszámot jelel az illető a decoder kontextusban számok fognak megjelenni, így amikor bizonytalan jelhez érünk a kontextus a szám felé billenti a kimeneti tokenek valószínűségét. Ez az információ mellesleg az encoder kontextusból is kinyerhető lenne közvetlenül, de a lépésekre bontás, egyszerű és jól működő módja a változó hosszúságú kimenet előállításának, ebben az esetben viszont, ha csak az encoder kontextus lenne a bemenet a decoder nem tudná, hogy a kimenet melyik tokenjét generálja. Szükség van tehát a decoder kontextusra is, az encoder kontextus feldolgozási lépéseinek számontartására.

### Saját modell kialakítása

A modell kimenete, és a decoder kontextus is szöveg, ezt viszont nem lehet közvetlenül a modell bemenetére helyezni. Az előző alpontban felvezetett módon tokenekre kell bontani, és a tokenekhez számot rendelni, például kategória index. A feladat jellegéből adódóan érdemes az egyes karaktereket tokennek választani. Egy gyakori lépés még az természetes nyelvfeldolgozás terén (NLP), hogy a modell megtanul több dimenziós vektorokkal reprezentálni tokeneket. Ez azért hasznos, mert a modell képes olyan vektor reprezentációt rendelni a tokenekhez, melyek hasonló kategóriák esetén valamilyen távolságmetrika szerint közel lesznek egymáshoz, nagyon különböző kategóriák esetén pedig pont ellenkezőleg távol esnek majd egymástól a vektortérben. Például ha szavak lennének a tokenek, a „macska” szó vektorreprezentációja tanulás után közel lenne a „cica” szó-hoz tartozó vektorhoz, viszont távol az „egér”-hez tratozótól. A konkrét feladatban az „a” karakterhez tartozó reprezentáció távol kerülhet a „b”-hez tartozótól, annak ellenére hogy a kategória indexeléskor sorban kiosztott számok tekintetében egymás mellett lennének, de jeleik nem hasonlítanak.

A póz adatokat nem lehet ilyen formában reprezentálni, de érdemes ennek mintájára olyan tanulható transzformációt végezni rajta, ami esélyt ad hasonló szemantikus reprezentáció kialakítására. Erre a korábbi statikus ujjbetűzés modellek című fejezetben felvezetett módon konvolúciós rétegeket alkalmaztam. Ezen rétegek az idő dimenzió mentén mintavételezik az egyes kulcspont típusokat.

#### Visszacsatolt neurális hálózatok

Egy lépés maradt hátra, mégpedig az encoder és decoder belső kialakítása. Első megközelítésben visszacsatolt (rekurrens - RNN) háló architektúrákat próbáltam ki. Ezeket szekvenciális adatok hatékony feldolgozásához fejlesztették ki. A "state" vagy "állapot" az RNN és hasonló visszacsatolt architektúrák kulcsfogalma. Egy RNN-nél az állapot egy belső memória, amely információt tárol az eddig feldolgozott adatszekvenciáról. Minden egyes lépésnél az állapot frissül az aktuális bemenet alapján, és ez az információ továbbítódik a következő lépésbe. Ezen keresztül az RNN képes "emlékezni" a korábbi lépésekre és figyelembe venni a múltbeli információkat a jelenlegi döntések során. Előny, hogy ugyanaz az RNN képes feldolgozni különböző hosszúságú szekvenciákat, de a hosszú távú függőségek tanulása nehézkes lehet az egyszerű RNN modellek számára. Három féle verziót próbáltam ki:

1. A SimpleRNN, ahogyan a neve is sugallja, az RNN modellek egyszerű változata. A fő jellemzője, hogy képes "visszaemlékezni" az előző lépések információira. Minden lépésnél az aktuális bemenet és az előző lépésből származó információ alapján döntést hoz. Ennek ellenére a SimpleRNN nehezen tanulja meg a hosszabb távú függőségeket az adatokban, ami korlátozza a felhasználhatóságát bizonyos feladatokra. Hosszú szekvenciák esetén gyakran előfordul, hogy tanulás során (backpropagation aloritmus), a korrekciós értékek (deriváltak) matematikailag túl nagyra, vagy éppen túl kicsire változnak, így nem tanul a hálózat. A többi RNN ezt próbálja gátolni.
2. GRU (Kapuzott Rekurziós Egységek) A GRU egy modernizált változata az RNN-nek, amely két "kaput" használ a döntéshozatalban: az újraindítási és a frissítési kaput. Ezek a kapuk lehetővé teszik a modell számára, hogy hatékonyabban döntsön arról, mely információkat kell megőrizni vagy eldobni az idő során. Ez segít a modellnek abban, hogy jobban kezelje a hosszabb távú információs függőségeket az adatokban.
3. LSTM (Hosszú Rövid Távú Memória) Az LSTM az RNN egy másik fejlett változata, amely még hatékonyabban kezeli a hosszú távú információs függőségeket. Három kaput használ a döntéshozatalban: a felejtési, a bemeneti és a kimeneti kaput. Az LSTM képes dönteni arról, mely információkat kell megőrizni, frissíteni vagy eldobni, ami lehetővé teszi számára, hogy még bonyolultabb szekvenciális mintázatokat is felismerjen és tanuljon.

A saját modellben az encoder-ben elhelyezett visszacsatolt réteg szekvenciálisan feldolgozza a bemenetet. A réteg kialakult belső állapotával inicializálva a decoder RNN rétegét, annak információja lesz az encoder kontextusáról. A decoder bemenetére pedig a kimenet egyre hosszabb részletei kerülnek. További lehetőség van arra is, hogy több réteget helyezzünk el láncolva egymás után.

#### Transzformer modellek

A transzformerek a mély tanulásban egy forradalmi architektúra típus, amely az "Attention is All You Need" című cikkben jelent meg 2017-ben, és melyet eredetileg szöveges adatok feldolgozására terveztek. Lényege az úgynevezett figyelem mechanizmusban (attention) rejlik, ami lehetővé teszi a modell számára, hogy az egész adatszekvenciára súlyozottan figyeljen, nem csak az aktuális vagy közvetlenül előző lépésekre, továbbá több párhuzamos „fejet” is tartalmazhat a figyelmi mechanizmus (multihead attention), ilyenkor különböző aspektusait tanulja meg megfigyelni az adatnak.

A transzformerek nem rendelkeznek beépített módszerrel a szekvenciális bemenetek sorrendjének figyelembevételére, ellentétben az RNN-ekkel, amelyek szekvenciálisan dolgozzák fel a bemeneteket. Emiatt van szükség a pozícionális kódolásra: hogy információt adjon a modellnek az egyes elemek pozíciójáról a szekvenciában. Ezt a kódolást úgy adják hozzá a bemeneti vektorokhoz, hogy az ne torzítsa az eredeti adatokat, de mégis információt adjon a modellnek az elemek sorrendjéről.

A transzformerek nem csak encoder-decoder formában léteznek, létezik még úgynevezett encoder only (BERT), és decoder only (GPT) változata is, ezek arra épülnek, hogy ki lehet úgy alakítani mind az encoder, mind a decodert, hogy egymás után kapcsolható legyen. A tanulmányban vizsgált modellt az eredeti 2017-es tanulmány alapján alakítottam ki. Ebben az encoder és decoder figyelmi mechanizmust alkalmaz a bemenetére, valamint a decoder az encodertől kapott kontextus vektorra. Figyelmi mechanizmusként multihead attention-t használtam, 4 fejjel egységesen, valamint 2 encodert és 4 decodert találtam optimálisnak a feladatra.

### Tanítás

#### Teacher forcing

Amikor soros adatokon alapuló modelleket tanítunk (például szövegfelismerés vagy fordítás esetén), az eredeti megközelítés az, hogy a modell által az előző lépésben előállított kimenetet használjuk fel bemenetként a következő lépés predikciójához. A gond ezzel a megközelítéssel az, hogy a tanítás során a modell hibája összeadódhat, mivel minden lépésbeli hiba hatással van az összes következő lépésre. A "teacher forcing" technika itt jön képbe. Az ötlet az, hogy a tanítás során a valódi célbemenetet használjuk fel bemenetként a következő lépés predikciójához, nem pedig a modell által az előző lépésben előállított kimenetet. Így a modell gyorsabban és hatékonyabban tanulhat.

Példaként vegyünk egy egyszerű fordítási feladatot, ahol a bemeneti mondat: "I am learning" és a célmondat (fordítás): "Tanulok". Ha nem használunk "teacher forcing"-ot, és a modell az első szó után nem megfelelő predikciót ad (például "Én"), akkor ez a hiba befolyásolhatja a következő predikciókat is. "Teacher forcing" használatával azonban a második szó predikciójához közvetlenül a "Tanulok" első szavát, azaz "Tanulok"-ot használjuk bemenetként, függetlenül attól, hogy a modell milyen predikciót adott az első szóra.

#### Metrikák és egyéb hiperparaméterek

A bemeneti adatok, mint póz szekvenciák, illetve hozzá tartozó szövegek nem egyforma hosszúságúak, viszont tanítás során nem egyesével kerülnek feldolgozásra, hanem csoportosítva (batch). A szekvenciák mindkét esetben ki vannak egészítve egy maximális méretre, hogy méreteik mind egyformák legyenek. Ezért a tanítás során használt hibafüggvénynek, illetve találati arány (accuracy) függvénynek is egy maszkolt verzióját használtam, amik ignorálják a kitöltő (padding) karaktereket.

Optimalizálóként az „adam” algoritmust alkalmaztam testreszabott tanulási rátával (learning rate). Az alkalmazott módszer a „warmup and decay” algoritmust követi. A tanulás elején, amikor a modell súlyai teljesen véletlenszerűek, túl nagy lépések a súlyok frissítésében instabilitást okozhatnak. A warmup periódus alacsony tanulási rátával kezdődik, így a modell lassabban alkalmazkodik a kezdeti időszakban, megelőzve a divergenciát. A tanulási rátát fokozatosan csökken (decay), a tanulás későbbi szakaszaiban megakadályozva, hogy a modell túlságosan alkalmazkodjon a tanító adathalmazhoz. (overfitting) Az RNN hálózatoknál „linear warmup cosine decay”-t, a transzformereknél pedig az „Attention is all you need” tanulmányban ajánlott „linear warmup inverse squareroot” formulát használtam.

eredmények

## Kiegészítő algoritmusok

Asd

## Értékelés

Asd

# Hagyomás jelelés

Asd

# Saját applikáció

How to save model properly

Preprocess included

Info function for mediapipe and write a function that uses that to gather coords

Write about whole tensorflow apparate

Put the testing code into a .py and call it app

Do the same for static fingerspelling maybe???

# Általános jelnyelv fordítása, kitekintés

Asd

# Köszönetnyilvánítás

A kutatás az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében.

# Irodalomjegyzék

Az Irodalomjegyzékben szereplő hivatkozásokat **Irodalomjegyzék sor** stílussal formázzuk, a címüket pedig **Irodalomjegyzék forrás** stílussal emeljük ki.

A szövegbe a hivatkozásokat a Kereszthivatkozás beszúrása (Insert cross-reference) funkcióval helyezzük el (példa egy így beszúrt hivatkozásra: [1]), így azok automatikusan frissülnek a hivatkozások átrendezésekor.

Irodalomjegyzék

1. Levendovszky, J., Jereb, L., Elek, Zs., Vesztergombi, Gy.: Adaptive statistical algorithms in network reliability analysis, Performance Evaluation - Elsevier, Vol. 48, 2002, pp. 225-236
2. National Istruments: LabVIEW grafikus fejlesztői környezet leírása, <http://www.ni.com/> (2010. nov.)
3. Fowler, M.: UML Distilled, 3rd edition, ISBN 0-321-19368-7, Addison-Wesley, 2004
4. Wikipedia: Evaluation strategy, <http://en.wikipedia.org/wiki/Evaluation_strategy> (revision 18:11, 31 July 2012)

Függelék