Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Biometrikus felhasználó azonosítás

Diplomaterv

Készítette
Boér Lehel

Konzulens
dr. Zainkó Csaba

2019. április 9.

Tartalomjegyzék

HALLGATÓI NYILATKOZAT

Alulírott *Boér Lehel*, szigorló hallgató kijelentem, hogy ezt a diplomatervet meg nem engedett segítség nélkül, saját magam készítettem, csak a megadott forrásokat (szakirodalom, eszközök stb.) használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Hozzájárulok, hogy a jelen munkám alapadatait (szerző(k), cím, angol és magyar nyelvű tartalmi kivonat, készítés éve, konzulens(ek) neve) a BME VIK nyilvánosan hozzáférhető elektronikus formában, a munka teljes szövegét pedig az egyetem belső hálózatán keresztül (vagy autentikált felhasználók számára) közzétegye. Kijelentem, hogy a benyújtott munka és annak elektronikus verziója megegyezik. Dékáni engedéllyel titkosított diplomatervek esetén a dolgozat szövege csak 3 év eltelte után válik hozzáférhetővé.

Budapest, 2019. április 9.

Boér Lehel
hallgató

Kivonat

Jelen dokumentum egy diplomaterv sablon, amely formai keretet ad a BME Villamosmérnöki és Informatikai Karán végző hallgatók által elkészítendő szakdolgozatnak és diplomatervnek. A sablon használata opcionális. Ez a sablon LaTeX alapú, a *TeXLive* TeX-implementációval és a PDF-LaTeX fordítóval működőképes.

Abstract

This document is a LaTeX-based skeleton for BSc/MSc theses of students at the Electrical Engineering and Informatics Faculty, Budapest University of Technology and Economics. The usage of this skeleton is optional. It has been tested with the *TeXLive* TeX implementation, and it requires the PDF-LaTeX compiler.

1. fejezet

Háttérismeretek

1.1. Biometrikus azonosítás

A biometria az emberek fizikai jellemzőinek mérésével és elemzésével foglalkozik. Alkalmazását tekintve három területet különböztetünk meg:

- Felhasználó ellenőrzés: Az azonosító rendszer a biometrikus adatot egy, korábban vetthez hasonlítja. Ez alapján dönt, hogy a felhasználó hozzáférhet-e a kívánt erőforráshoz. Ilyen egy ujjlenyomat-olvasóval ellátott mobiltelefonon a képernyőzár feloldása. A felhasználó ellenőrzés arra ad választ, hogy az illető az-e akinek mondja magát.
- Felhasználó azonosítás: Az azonosító rendszer a biometrikus adatot több korábban vett mintához hasonlítja és arra ad választ, hogy ki a felhasználó; azaz beletartozike a korábban eltárolt biometrikus adatokból álló csoportba vagy nem. Ilyen lehet például egy ujjlenyomat-leolvasóval ellátott beléptetőrendszer cégek esetében.
- Duplikátum detektálás: Annak ellenőrzése, hogy egy felhasználó egynél többször szerepel-e egy adatbázisban. Csalások, például szociális támogatást többször igénylők kiszűrésére használják.

Az első biometrikus azonosítási eljárás az ujjlenyomatvételen alapuló személyiségazonosítás volt, amely a modern kriminalisztika világában terjedt el, de manapság már megtalálható okostelefonokban, biometrikus beléptetőrendszerekben is.

A biometrikus azonosítást az ún. biometrikus azonosító rendszer végzi el. A folyamat során a biometrikus azonosító rendszer mintát vesz az azonosítandó egyén egy vagy több előre meghatározott fizikai jellemzőjéről, és ezekről digitális lenyomatokat képez. Az első, regisztrációs fázisban a biometrikus minta lenyomatát a rendszer egy adatbázisban eltárolja, majd később az azonosítás során az aktuális mintát összeveti a korábban rögzítettel és dönt az egyezésről. Ahhoz, hogy az ember egy fizikai jellemzőjét biometrikus adatként használhassuk, a következő elvárásokat támasztjuk vele szemben:

- Általánosság: A biometrikus adattal minden egyénnek rendelkeznie kell.
- Egyediség: A biometrikus adatnak egyedinek kell lennie a releváns populáción belül.
- Állandóság: A biometrikus adat nem, vagy csak keveset változzon az idő eltetével.
- Mérhetőség: Az biometrikus adat az egyén részéről legyen könnyen mérhető testi adottság.

- Teljesítmény: A biometrikus azonosító rendszerek teljesítménye: gyorsaság, pontosság, technológia.
- Elfogadottság: A releváns populáción belül a mérési eljárás mennyire elfogadott (emberi méltóság megőrzése).
- Biztonság: Mennyire nehéz utánozni, hamisítani a biometrikus adatot?

A biometrikus adat lehet fiziológiai (DNS, arc, ujjlenyomat, írisz) vagy viselkedési (hang, írás, gesztusok). Mivel ezek az adatok statisztikai jellegűek, megbízhatóságuk változó. Minél több adat van egy mintában, annál egyedibb, és minél nagyobb a releváns populáció (eltárolt minták összessége), annál valószínűbb, hogy találunk két hasonló mintát. Ennek elkerülésére manapság terjednek a multimódusú biometrikus azonosító rendszerek, amelyek több biometrikus adatot felhasználva végzik ez az ellenőrzés, azonosítás és duplikátum detektálás feladatát.

1.2. Beszélőfelismerés

Az emberi kommunikáció során fontos feladat a beszélő partner felismerése. A telekommunikációs technológia fejlődése miatt elterjedt a telefonon vagy interneten történő hangalapú kommunikáció; a telefonos felhasználófelismerés mint biometrikus azonosítási módszer megjelent már banki alkalmazásokban, call centerekben és az elektronikus kereskedelemben is (mobiltelefonos vásárlás). Az elektronikus kommunikáció során sokszor csak a beszélő hangjára hagyatkozhatunk, az alapján ismerhetjük fel az illetőt. A beszélőfelismerést háromféle módon végezhetjük:

- Naiv beszélőfelismerés: Az emberi, naiv beszélőfelismerés során az ismerős hangokat meglepően nagy pontossággal ismerjük fel.
- Törvényszéki beszélőfelismerés: A törvényszéki szakértői vizsgálat eredménye.
- Automatikus beszélőfelismerés: A beszélőfelismerést számítógépes rendszer végzi.

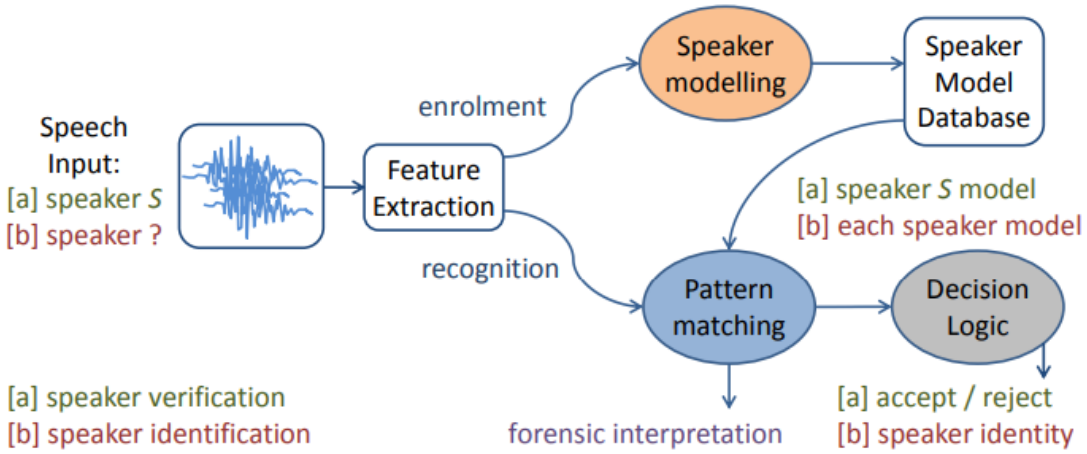
A beszélőfelismerés alatt három szűkebb fogalmat értünk. Ha a folyamat során az ismeretlen beszélőről azt ellenőrizzük, hogy az-e akinek állítja magát beszélő ellenőrzésről van szó. Beszélő szegmentáláskor a hangmintát homogén csoportokra bontjuk a beszélő személye alapján. Végül beszélő azonosításról beszélünk, ha az illető hangját rögzített hangok egy csoportjával vetjük össze és azt szeretnénk eldönteni, hogy melyikhez hasonlít a legjobban. Utóbbi felveti a kérdést, hogy mi történik ha a beszélő nem tagja a csoportnak.

Emiatt megkülönböztetjük a nyitott és zárt halmazú beszélőazonosítást. Utóbbi esetén csak olyan beszélőket ismerünk fel, akikről van hangminta az adatbázisban, míg az előbbinél ismeretlen beszélők is megjelenhetnek, így ezt is kezelni kell.

A beszélőfelismerés továbbá lehet szöveg-függő és szöveg-független attól függően, hogy a felismerő rendszer egy előre meghatározott mondatot vár, vagy bármilyen hangminta alapján működik.

1.3. Az automatikus beszélőfelismerés története

Az automatikus beszélőfelismerést egy számítógépes program végzi emberi beavatkozás nélkül. Az első automatikus beszédfelismerő rendszert a Texas Instruments fejlesztetése volt és 1977-ben publikálták. A rendszer szövegfüggő beszélőellenőrzésre volt képes és az évek során a téves elutasítási és elfogadási rátája 1% alatt maradt. A hetvenes évek óta a a



1.1. ábra. Automatikus beszélőfelismerő rendszer architektúrája.

beszélőazonosító és ellenőrző rendszerek rengeteget fejlődtek kezdve a vektor kvantálástól kezdve a GMM modelleken át a mély neurális hálókig.

Az automatikus beszélőfelismerő általános működését az alábbi (!) ábra szemlélteti. A beszélő-ellenőrzés és beszélőazonosítás során is az első lépés rögzített *jellemzők kinyerése*. Ezután az első, *tanító fázisban* referencia modelleket készítünk az egyes beszélőkhöz a jellemzők alapján, amelyeket eltároljuk a beszélő adatbázisban. Beszélő-ellenőrzés esetén ebben a fázisban egy küszöbérték is meghatározásra kerül. *Teszt fázisban* a rendszer kinyeri ugyanazokat a jellemzőket az aktuális hangmintából, majd *jellemző összehasonlítás* történik.

1. beszélő-ellenőrzés esetén megkeresi az ellenőrizendő személyhez tartozó modellt az adatbázisban és összehasonlítja az aktuális jellemzőkkel. Ha az eredmény a küszöbszint felett van, a rendszer egyezést mutat.

2. beszélőazonosítás esetén az aktuális jellemzőket az összes modellel összehasonlítja, majd a legjobb egyezés mellett dönt.

1.3.1. Jellemző kinyerés

A jellemző kinyerés célja a dimenzió csökkentése és a beszélőspecifikus információk kinyerése. Mivel a beszéd komplex jel, a beszélő azonosítása szempontjából felesleges információkat is hordoz. Ilyen például a környezet és a csatorna zaja. A kinyert jellemzőket a hangminta terjedelme alapján osztályozzuk. *Rövidtávú jellemzők* a 20-30 ms-os keretekből kinyert mel-frekvenciás és lineáris prediktív kepsztrális együtthatók (MFCC és LPCC). A *prozódikus jellemzők* kinyerése 100 ms-os terjedelemben történik és a beszéd ritmusát, a hangmagasságot és a sebességet jellemzik. A hosszútávú jellemzőket a jel akár perc hosszú kereteiből nyerjük ki. Ezek képesek reprezentálni a beszélő akcentusát illetve a szavak szemantikáját és az idiolektust.

A beszélőfelismerő rendszerek teljesítményét javította, ha a jellemzőket csak a hangminta azon részeiből nyerték ki, amikben beszéd is volt. Erre alkalmazott technika a *Voice Activation Detection* (*VAD*).

1.3.2. Jellemző normalizálás

Jellemzőkinyerés során próbáljuk kiszűrni a beszélő szempontjából értékes részéket, ugyanakkor nincs tökéletes jellemző, amely ne változna a környezet hatására. Ezt a változást segítik minimalizálni a normalizálási módszerek.

1.3.3. Beszélő modellek

Kezdetben a vektor kvantálás volt az elterjed modellezési módszer, amit később a *Gaussian Mixture Model* (*GMM*) váltott fel. A GMM egy adathalmazt több normális eloszlás keverékeként ír le és képes nem felügyelt módon klaszterezni az adatokat. Egy beszélőhöz egy valószínűségi sűrűségfüggvényt rendel, amely különböző pontokban kiértékelve (például teszt fázisban a beszélőtől kinyert jellemzők) egy valószínűséget ad a két beszélő hasonlóságára.

A GMM megközelítés főleg beszélőazonosításra alkalmas. Beszélő-ellenőrzéshez szükség volt egy másik modellre is, ami képes leírni minden más beszélőt az ellenőrizendőn kívül. Erre adott megoldást az *Universal Background Model* (*UBM*). Később jobb teljesítményt értek el, ha a teszt fázisban a beszélőkkel először UBM modelleket tanítottak és ezekből származtattak GMM-eket. Ezt nevezik GMM-UBM módszernek.

Mivel a tanító és teszt hangminták eltérő hosszúságúak lehetnek, szükség volt egy fix hosszúságú reprezentációra, ezt oldották meg a GMM szupervektorok, amelyeket az akkori megközelítés szerint szupport-vektor gépekkel vagy faktoranalízissel használtak.

Az utóbbi két módszer előnyeit kombinálva megszületett az $i-vektorok$, amelyet követve eljutunk a mai state-of-the-art módszerhez, a mély neurális hálózatokhoz (DNN).

1.4. Korábbi eredmények

Szerző (év)	**Szervezet**	**Adatbázis**	**Módszer**	**Jellemzők**	**Hang típusa**	**Pontosság**
Douglas A. Reynolds (1995)	Lincoln Laboratory	49	MFCC	rövid kifejezések	telefon	96.8 %
Rabah W. (2004)	King Abdulaziz University	20	SVD-alapú algoritmus	LPC/Cepstral	iroda	94 %
Yang Shao (2008)	Ohio State University	34	GFCCs	hallási jellemzők	telefon	~99.33 %
P. Krishnamoorthy (2011)	TIMIT	100	GMM-UBM	MFCC	labor	80 %
Alfredo Maesa (2012)	Voxforge.org	250	MFCC	spektrális jellemzők	beszéd-adatbázis	$>$ 96 %
Sharada V. Chougule (2015)	Finolex Academy of Management and Technology	97	NDSF	spektrális	labor	~98-100 %

1.1. táblázat. Korábbi eredmények szövegfüggetlen beszélőazonosítás terén.

2. fejezet

Beszélőazonosító rendszerek napjainkban

A fejezet bemutatja a tanítás során használt beszédadatbázisokat, két neurális hálózat alapú, zárt-halmazú, automatikus beszélőfelismerő rendszert; a *WaveNet classifiert* és a *SincNetet* illetve az ezekkel elért eredményeket.

2.1. Beszédadatbázisok

2.1.1. TIMIT

A TIMIT beszédkorpuszt automatikus beszédfelismerő rendszerek fejlesztéséhez tervezték. 630 beszélőtől tartalmaz mintákat amerikai angol nyelven a 8 legelterjedtebb nyelvjárásban. A TIMIT archívum tartalmaz egy TRAIN és egy TEST mappát, ezek tanításhoz és teszteléshez valók. Ezeken belül további, a dialektusok sorszámával (DR1, ..., DR8), azon belül a beszélő azonosítójával elnevezett könyvtárak találhatók. Egy beszélőhöz 10 db beszédminta tartozik 16 kHz-es NIST SPHERE fájlok formájában.

Dialektus regió (DR)	**Férfi**	**Nő**	**Összesen**
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)

2.1. táblázat. A beszélők eloszlása dialektusok szerint.

A dialektus régiók a következők:

- DR1: New England
- DR2: Northern
- DR3: North Midland
- DR4: South Midland
- DR5: Southern
- DR6: New York City
- DR7: Western
- DR8: Army Brat (moved around)

A NIST SPHERE formátum az elején definiál egy fix hosszú fejlécet, amit a hang bináris kódolása követ. Erre figyelni kell a hangfájlok beolvasásánál, Python esetében nem minden hangfeldolgozó könyvtár támogatja. Ilyen esetben kézzel el kell távolítani a fejlécet a fájlok elejéről. A hangfájlokhoz tartozik egy szöveges dokumentum ami az elhangzott szöveget és annak a wav fájlbeli helyét tartalmazza. Továbbá egy WRD fájl írja le a szavakat és egy PHN kiterjesztésű a fonémákat, illetve azok időbeni elhelyezkedését a wav fájlban.

A hangfájlokhoz tartozó egyéb fájlok beszédfelismerés szempontjából fontosak. Mivel én a beszédkorpuszt beszélőfelismerésre használtam, csak a hangfájlokra volt szükségem. A TEST mappában - mivel a TIMIT-et alapvetően beszédfelismeréshez tervezték - teljesen különböző beszélők vannak a TRAIN mappához képest, ezért a TRAIN mappabeli beszélőket kell felosztani tanításhoz és teszteléshez.

2.1.2. CMU Arctic

A CMU Arctic beszédadatbázist beszédszintézis kutatásokhoz tervezték. 18 adathalmazt tartalmaz 18 különböző embertől más-más akcentussal, angol nyelven. Egy emberhez több száz beszédminta tartozik wav fájlok formájában.

2.1.3. Adatok előfeldolgozása

Az előfeldolgozó szkript a TIMIT adatbázis esetében a hangmintákról eltávolítja a NIST Sphere fejlécet és a mondat előtti és utáni szüneteket. Ezután normalizálja a hangmintákat. A CMU Arctic esetében a hangminták alapból normalizálva vannak, ezért csak azonos méretűre kell vágni őket az egységes bemeneti dimenziók érdekében (ahogy a TIMIT esetében is).

2.2. Mérési elrendezés

Google Colab...

2.3. WaveNet classifier

A *WaveNet classifier* egy módosított WaveNet architektúra beszélőidentifikációhoz.

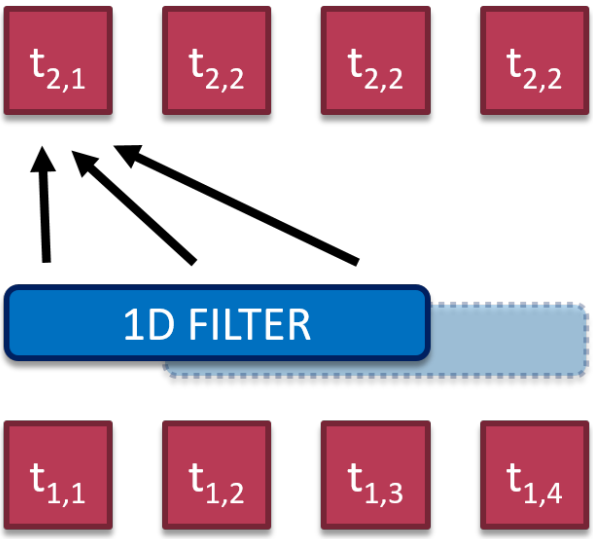
2.3.1. WaveNet

A WaveNet egy mély neurális hálózat audio hullámformák generálásához, amelyet a Google DeepMind publikált 2016-ban. Az ötletet az akkori felfedezések adták neurális autoregresszív generatív modellezés terén, amelyeket komplex eloszlások, például képek modellezésére használtak (van den Oord et al., 2016a;b). Ezt felhasználva audio hullámformák generálásában új eredményeket értek el.

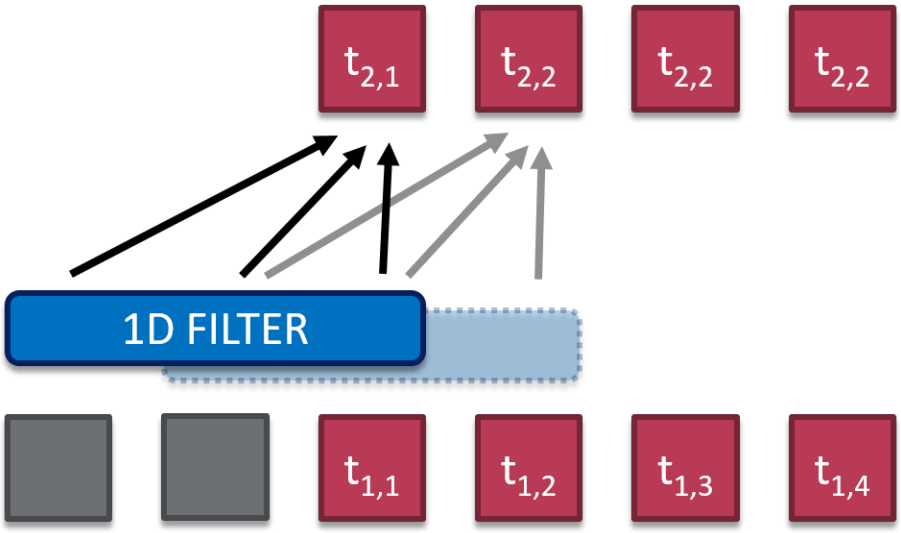
- Képes olyan természetes hangzású beszéd hullámformák generálására, amit korábban parametrikus vagy konkatenatív beszédszintézissel sosem értek el.
- Az audio hullámformák generáláshoz szükséges nagy receptív mezőt hatékonyan, nyújtott kauzális konvolúciókkal implementálja.
- Ha a modellt a beszélők identitásával tanítják, képes új hangok generálására.
- A zene generálás és a beszédfelismerés terén is ígéretesnek bizonyult.

2.3.1.1. Nyújtott kauzális konvolúció

A modell autoregresszív, vagyis a kimenete korábbi időpillanatokban felvett értékeitől függ. Ez azért fontos, mert a WaveNet egy generatív modell. A generált hullámforma t-edik időpillanatbeli értéke nem függhet jövőbeli értékektől. A kauzalitás legegyszerűbb implementációja ha legalább a *kernel-1* méretű paddinget adunk a konvolúcióhoz.



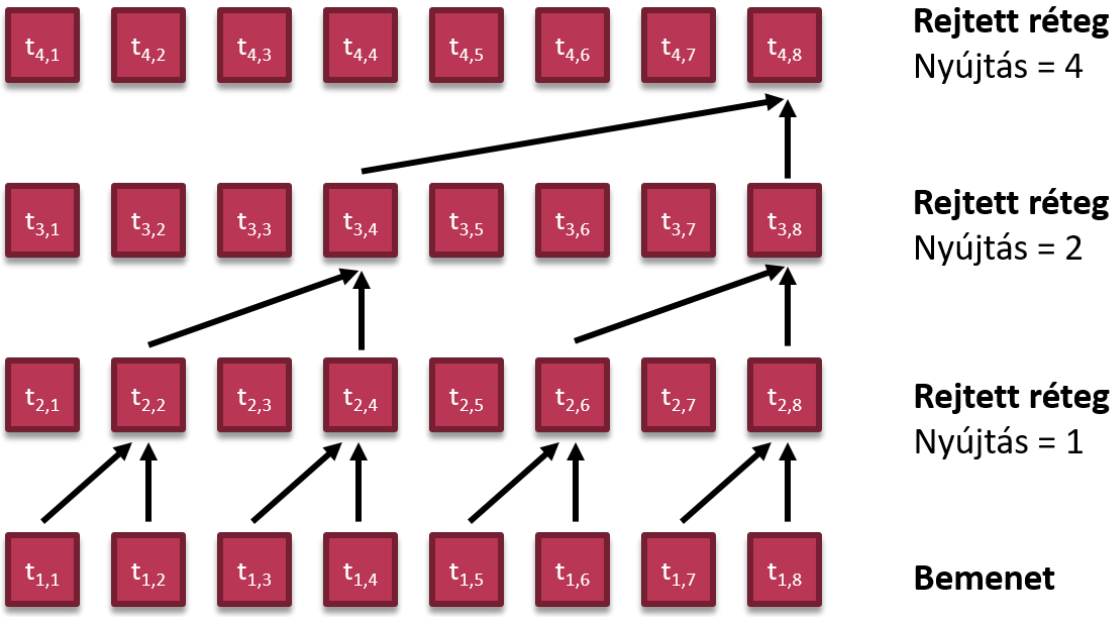
2.1. ábra. 1D nem kauzális konvolúció.



2.2. ábra. 1D kauzális konvolúció paddinggel.

A 2.1 ábra egy nem kauzális 1D konvolúciót mutat. A $t_{i,j}$ az i-edik rétegbeli j-edik neuron. Látható, hogy a $t_{2,1}$ jövőbeli időpillanatokból kap értékeket a szűrőn keresztül. Ennek megoldása a szűrő eltolása padding segítségéve. Ezt szemlélteti a **??** ábra, ahol az egyes neuronok csak korábbi időpillanatokból kapnak értékeket.

A beszéd generálásánál a t-edik időpillanatban a hullámforma értéke a korábbi adatoktól függ. Ahhoz, hogy magas frekvenciájú, pl. 16 kHz frekvencián mintavételezett hangadattal tanítsuk a hálózatot nagy receptív mezőre van szükség. A receptív mező az a szélesség, amit a szűrő lát a bemenetből. 16 kHz esetén egy másodpercnyi jelet 16000 szám reprezentál. Ahhoz, hogy a hálózat helyesen jósolja meg a következő generált értéket, a hosszú távú dependenciákat figyelembe kell vennie, tehát a receptív mező méretét elég nagyra kell megválasztani. A probléma ekkora mezők esetén, hogy sok konvolúciós réteget igényelnek (egy korábbi időpillanatbeli adat plusz egy konvolúciós réteget igényel), ami növeli a számítási komplexitást. A nyújtott konvolúciók erre adnak hatékony megoldást. A filter meghatározott távolságokkal kihagy valamennyi inputot, majd figyelembe vesz egyet. Egymás utáni rétegekben a nyújtási tényezőt exponenciálisan növelve a receptív mező is exponenciálisan fog nőni.



2.3. ábra. 1D nyújtott kauzális konvolúciós rétegek.

2.3.1.2. SoftMax eloszlás

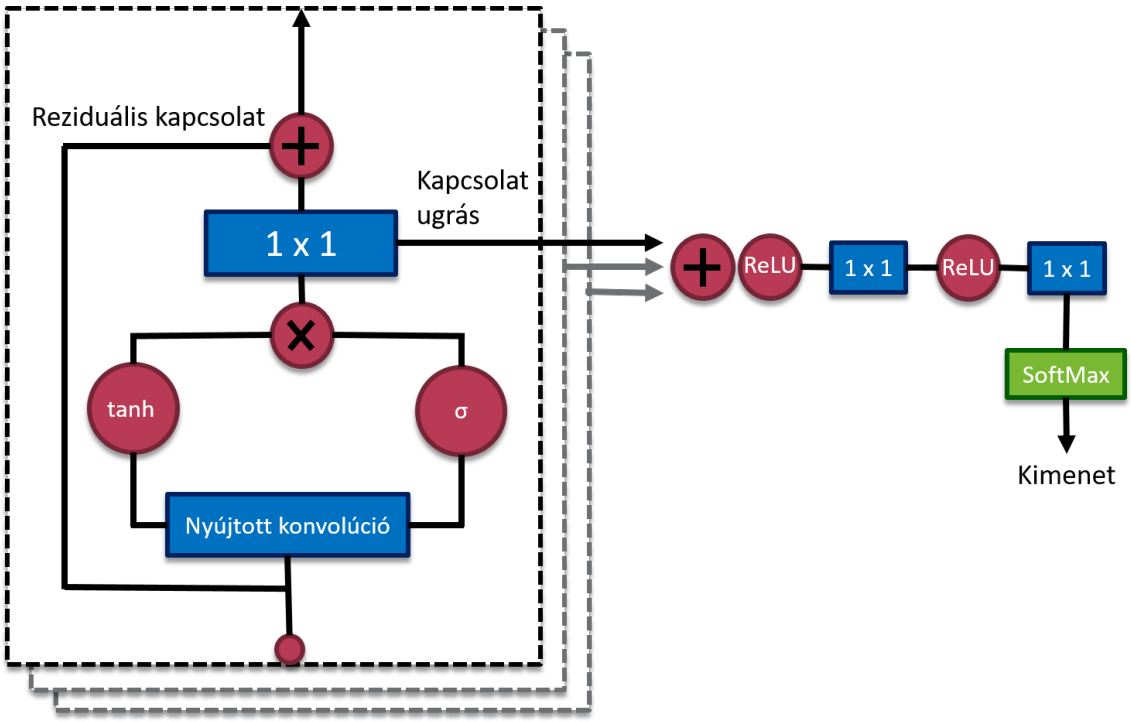
A WaveNet SoftMax réteget használ a $p(x_t|x_1, \ldots, x_t)$ feltételes valószínűségi eloszlás modellezésére. Az audio jeleket általában 16 bites egészekkel kódolják, amelyek 65536 értéket vehetnek fel. Ebben az esetben a SoftMax rétegnek 65536 valószínűséget kell kimenetként adnia, melyek összege 1. A μ-law kvantálást alkalmazva a beszédjel 256 biten kódolható és később az inverz transzformációval jó minőségben visszaállítható.

Az emberi hallás sokkal érzékenyebb alacsony amplitúdójú hangok kvantálási zajára, mint a magasabbakéra. Erre alapozva a *mu*-law kvantáló a jelet egy logaritmikus függvénnyel kvantálja úgy, hogy az alacsonyabb amplitúdójú jelek nagyobb felbontással (több bittel), a magasabbak pedig kisebbel lesznek kódolva. Ez növeli a SoftMax réteg hatékonyságát is, mert nagyobbak lesznek a valószínűségek közötti különbségek.

A tanítást a WaveNet klasszifikációs problémaként kezeli. A bemeneteket OneHot kódolással adjuk meg, a SoftMax réteg pedig az így kódolt egészekre ad valószínűségi eloszlást.

2.3.1.3. Reziduális blokkok

A WaveNet architektúra egymáshoz csatolt reziduális blokkokból és ún. kapcsolat-ugrásokból (skip-connection) épül föl. A reziduális hálózatok előnye, hogy orvosolják az elenyésző gradiens problémát, így sokkal mélyebb hálózat építhető.



2.4. ábra. WaveNet architektúra.

Egy reziduális blokk bemenete egy 2x1-es konvolúciós rétegen megy keresztül. Balra egy *tanh*, jobbra egy szigmoid aktivációs függvényen haladnak át, majd elemenkénti szorzás és 1x1 konvolúció után egyrészt átugorja a reziduális kapcsolatot, illetve azzal együtt bemenetként szolgál a következő reziduális egységnek. Az 1x1 konvolúciós rétegek a dimenzionalitás változtatására szolgál. Külön 1x1 konvolúciós szűrők skálázzák a kimenetet a következő reziduális blokk bemenetére, és a kapcsolat-ugrásokhoz.

2.3.2. Módosított WaveNet architektúra

A módosított WaveNet architektúra segítségével a WaveNetet beszélőfelismerésre használhatjuk.

```
from WaveNetClassifier import WaveNetClassifier

wnc = WaveNetClassifier((96000,), (10,), kernel_size = 2, dilation_depth = 9,
                        n_filters = 40, task = 'classification')

wnc.fit(X_train, y_train, validation_data = (X_val, y_val), epochs = 100,
        batch_size = 32, optimizer='adam', save=True, save_dir='./')
```

```
y_pred = wnc.predict(X_test)
```

A WaveNetClassifier objektum paraméterei:

- *input_shape*: Bemeneti dimenziók tuple formájában. Például ha a bemenet egy 6 s hosszú hullámforma 16 kHz-en mintavételezve, a bemeneti dimenziók (96000,)
- *output_shape*: Kimeneti dimenziók tuple formájában. Például ha 100 osztály szerint klasszifikálunk, a kimeneti dimenziókból képzett tuple (100,).
- *kernel_size*: A konvolúciós filter/kernel mérete a reziduális blokkokban.
- *dilation_depth*: A reziduális blokkok száma.
- *n_filters*: A konvolúciós filterek száma a reziduális blokkokban.
- *task*: Klasszifikáció vagy regresszió.
- *regression_range*: A regresszió céltartománya lista vagy tuple formátumban.
- *load*: Előző WaveNetClassifier betöltése (bool).
- *load_dir*: A betölteni kívánt modell könyvtára.

2.3.3. Eredmények

A WaveNet classifiert mindkét beszédadatbázison teszteltem. A TIMIT beszédkorpusszal csak kevés beszélő esetén ért el jó eredményt, több mint 20 beszélő esetén a modell nem tanult. Ennek valószínűsített oka, hogy a TIMIT adatbázis beszélőnként 10 hangmintát tartalmaz.

Beszélők száma	18
Minta/beszélő	100
Minta össz.	1800
Minta hossza	4000
Epochok száma	43
Hiba	0.0013
Pontosság	1.0

2.2. táblázat. Paraméterek CMU Arctic adatbázissal.

!!! teszthalmaz mérete !!!
Tanítás után a teszt adathalmazon a hálózat 96.799 %-os pontosságot ért el.

2.4. SincNet

2.4.1. Eredmények

3. fejezet

Meta learning

Ebben a fejezetben ismertetem a metatanítás fogalmát, a megközelítéseket. Megmutatom, hogy mit old meg a few-shot learning és ez hogyan felhasználható nyílt halmazú beszélőfelismerő rendszerekben.

> "— Miért van szüksége a neurális hálózatoknak sok tanítómintára a jó teljesítéshez? Egy két éves gyerek miután látott pár autót képes felismerni azt. — Egy ekkora gyereknek volt két éve tapasztalatokat szerezni olyan dolgokról, amik nem autók voltak. Biztos vagyok benne, hogy ez fontos szerepet játszott a dologban."
>
> (Quora)

Az emberek képesek hasznosítani a korábban szerzett tudást. Ha valaki megtanult biciklizni, utána sokkal könnyebben motorbiciklire ül. Felismerünk dolgokat úgy, hogy előtte csak néhányszor láttuk élőben vagy csak képen láttuk.

Az emberekkel szemben a neurális hálózatokkal két probléma van.

1. Nem tanulnak effektíven, sok tanítómintát igényelnek egy feladat megtanulásához.
2. Nem hasznosítják a korábban, más feladatok által megszerzett tudást.

Metatanítás alatt olyan tanító módszerek összességét értjük, amelyek felhasználják a korábban szerzett tudást és hasznosítják azt a későbbi feladatok során. A metatanuló modellek általánosítják a tapasztalataikat és könnyen adaptálódnak új feladatokhoz. A könnyű adaptáció alatt kevés tanítómintával végzett tanítást értünk, más néven finomhangolást.

3.1. Few-shot learning

A neurális hálózatok rengeteg tanítómintát igényelnek. A kevés tanítómintával tanított hálózatok túltanulnak; a tanítóhalmazon jó eredményt mutatnak, de csökken az általánosító képességük, így a teszthalmazon jelentősen rosszabb eredményt érnek el. Ugyanakkor ha sok tanítómintával tanítjuk a hálózatokat, az adott feladatokra már jól általánosítanak, de magukra a feladatokra tanítjuk túl őket. További hasonló feladatokra nem tudnak általánosítani.

A few-shot learning azzal a problémával foglalkozik, hogy hogyan építhetünk olyan modelleket, amelyek képesek új feladatokat gyorsan megtanulni. A modellt sokféle feladatra tanítjuk, de minden feladathoz kevés tanítóminta tartozik. A tanítás után új feladat esetén a modell kevés tanítómintával képes jól megtanulni azt.

A few-shot learningre való képességet az n-way k-shot feladattal szokták mérni.

1. A modell kap egy eddig nem látott osztályból egy tesztmintát.
2. Kap továbbá egy ún. segéd adathalmazt, ami az összes k eddig nem látott osztályból n mintát tartalmaz.
3. Az algoritmus el kell döntse, hogy melyik osztályba tartozik a tesztminta a segédhalmazból.

A beszélőfelismerő rendszereket tekintve nagyon fontos, hogy a neurális hálózat rugalmas legyen a tanulás szempontjából. Tegyük fel, hogy egy cég telefonos, szövegfüggetlen, nyílt halmazú beszélőfelismerést szeretne abból a célból, hogy a betelefonáló kliensek adatait rövid hangminta után a rendszer automatikusan kilistázza az operátornak. Egy nagyobb telefontársaság esetén ez több ezer beszélőt is jelenthet. Hagyományos neurális hálózatokkal minden egyes klienstől több perces hangmintára lenne szükség, hogy ne tanítsuk túl a modellt.

A másik, még nagyobb probléma, hogy új, vagy távozó kliens esetén újra kell tanítani a teljes hálózatot. A few-shot learning tökéletesen megoldja mindkét problémát. Kevés tanítómintával – regisztrációkor pár kérdés kliensenként – működik a rendszer és rugalmas is tanítás szempontjából; új kliens esetén hozzáadjuk a segédhalmazhoz a hangmintáját, távozáskor pedig eltávolítjuk.

3.2. Metrikus metatanítás

Metrikus metatanítás során a modell a tanítóminták halmazán egy távolságfüggvényt tanul meg.

$$P_\theta(y|x, S) = \sum_{(x_i,y_i)\in S} k_\theta(x, x_i)y_i \tag{3.1}$$

A valószínűsége annak, hogy a θ modellparaméterekkel az x minta S segédhalmaz mellett az y osztályba tartozik egyenlő a segédhalmazba tartozó címkék súlyozott összegével. A súlyt a k_θ ún. kernel függvény számolja ki. A kernel függvény az a távolságfüggvény, amit a modell megtanul.

A megközelítésre több implementáció is létezik, mint például a Matching Network, Prototypical Network, Relation Network és a konvolúciós sziámi hálózat. Ezek közül az utolsót fogom részletesen bemutatni.

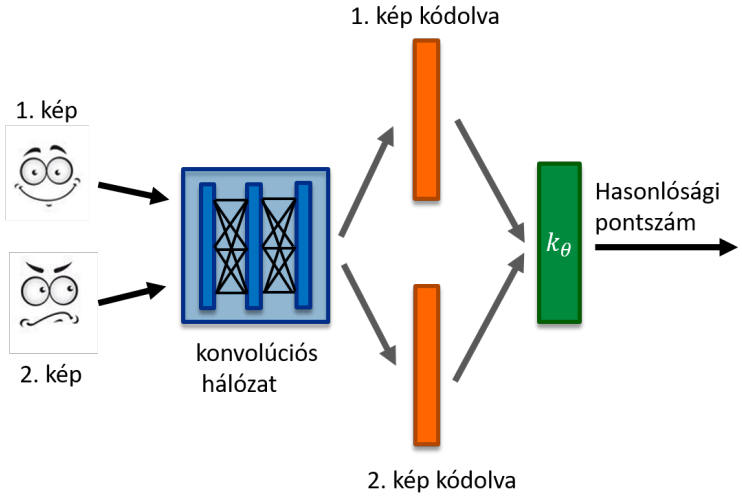
3.2.1. Konvolúciós sziámi hálózatok

A konvolúciós sziámi hálózatok két konvolúciós ikerhálózatból épülnek fel, amelyek megosztott súlyokat és rétegeket használnak. A hálózat két bementet kap, és miután a konvolúciós rétegek kiszámolták a jellemző vektorokat, ezeknek egy θ távolságfüggvénnyel méri a hasonlóságát. Ha a hasonlósági pont meghalad egy küszöbértéket, a két képet hasonlónak tekintjük, egyébként különbözőnek.

Sok ábrán a sziámi hálózatokat két azonos hálózattal reprezentálják. Valójában mivel a két ikerhálózat súlyai és rétegei megosztottak, elég egy hálózatot használni és csak a jellemző vektorokat eltárolni, így kevesebb erőforrást használunk.

Működésük közben egy mintáról nem azt tanulják meg, hogy melyik osztályba tartozik, hanem a különbséget a többi mintához képest. Tanítás közben a konvolúciós hálózatot súlyait javítják, hogy az általa képzett kódolások azonos minták esetén azonosak, különbözők esetén pedig különbözőek legyenek, így a távolságfüggvény jól fog működni.

A konvolúciós sziámi hálózatok megoldást jelentenek a few-shot learning problémára, mert kevés tanítómintával is jól működnek. Vegyük példaként egy vállalat arcfelismerő



3.1. ábra. Konvolúciós sziámi hálózat architektúrája.

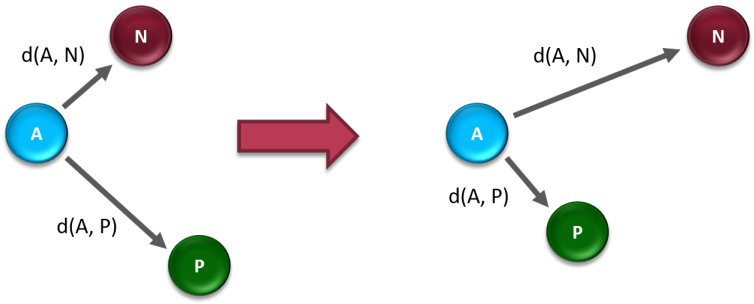
rendszerét. A sziámi hálózatnak elég egy segédhalmazban egy-egy képet eltárolni minden alkalmazottról. Amikor egy alkalmazott a rendszert használja, a képét összehasonlítja a segédhalmazban lévő képekkel és a hasonlóság alapján dönt. Ezzel két problémát old meg:

1. Egyrészt nem szükséges minden alkalmazottról több képet készíteni.
2. Másrészt új alkalmazottak érkezése, vagy egy alkalmazott távozása után nem szükséges újra tanítani a hálózatot.

A sziámi hálózatoknak a legismertebb költségfüggvénye a *triplet loss* függvény. Ezt a *gradient descent* módszerrel optimizálva tanítjuk a hálózatot. A triplet loss három bemenetet igényel, amelyek jelen példában a képekből képzett jellemző vektorok. Az egyik egy rögzített kép vektora, ez az ún. *anchor*. A másik kettő pedig egy pozitív és egy negatív minta. Az egyik ugyanabból az osztályból származik mint az *anchor* kép, a másik különbözőből.

$$\mathcal{L}_{triplet}(A, P, N) = \max(d(A, P) - d(A, N) + m, 0) \tag{3.2}$$

A d a távolságfüggvény, ami lehet például euklideszi távolság. A *triplet loss* veszi az anchor és a pozitív meg az anchor és a negatív minta távolságainak különbségét, majd ezt eltolja az m küszöbértékkel. Ha az előbbi pozitív ezt veszi eredményül, egyébként nullát.



3.2. ábra. A triplet loss függvény csökkenti a távolságot a hasonló és növeli a különböző minták között.

Akkor jó a vektorok elhelyezkedése a metrikus térben, ha a hasonlók között a távolság kicsi, a különbözők között pedig nagy. Azt szeretnénk elérni, hogy $d(A, P) \leq d(A, N)$ fenn

álljon. Átrendezve a $d(A,P) - d(A,N) \leq 0$ egyenletet kapjuk. Ezt kielégíti a $d(A,P) = 0$, $d(A,N) = 0$ megoldás. A másik triviális megoldás, ha a pozitív és negatív minta kódolása ugyanaz lenne, ekkor ugyanis $d(A,P) = d(A,N)$ így $d(A,P) - d(A,N) = 0$. Szeretnénk, ha a neurális hálózat nem nullvektorokkal vagy azonos vektorokkal kódolná az összes képet, ezért hozzáadunk egy m küszöbértéket az egyenlethez.

$$\begin{aligned} d(A,P) + m &\leq d(A,N) \\ d(A,P) - d(A,N) + m &\leq 0 \end{aligned} \tag{3.3}$$

Ideális esetben a $d(A,P) - d(A,N) + m$ negatív, ilyenkor a veszteség 0. Ha nem így van, a triplet loss ezt a veszteséget adja vissza. A költségfüggvény a tanítóhalmazban lévő tripletekre alkalmazott triplet lossok összege. Ezt minimalizálva a 3.2 ábrán látható távolságok csökkentése, növelése történik.

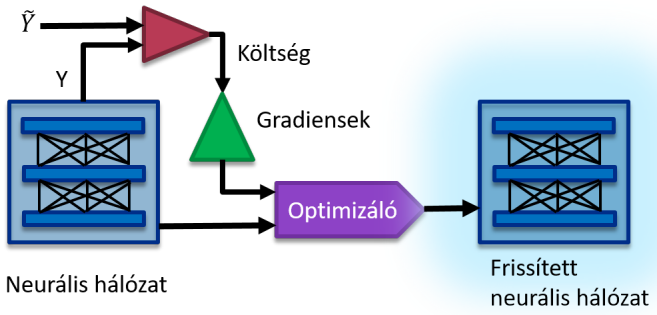
3.3. Optimizációs metatanítás

Az optimizációs algoritmusok mindig egy célfüggvényt szélsőértékét keresik. Neurális hálózatokban ez a célfüggvény a költségfüggvény felel meg és függ a modell tanulható paramétereitől. A optimizációs metatanítási megközelítések a modellek egyes paramétereit szintén modelleknek tekintik. Ezek a paraméterek tipikusan a modell kezdeti paraméterei (súlyok, eltolássúly) és az optimizációs algoritmus.

A fejezetben bemutatok egy megközelítést, ahol az optimizációs algoritmust modellezik az adott feladat tanításának felgyorsítása érdekében és két másik few-shot learning megoldást: a MAML és Reptile algoritmusokat, ahol a modell kezdeti paramétereit állítják be úgy, hogy könnyen adaptálható legyen új feladatokhoz.

3.3.1. Optimizáló algoritmus modellezése

Egy neurális hálózat tanulását mutatja a 3.3 ábra. A neurális hálózat által számolt kimenet és az elvárt kimenet közötti különbség alapján a költségfüggvénnyel számoljuk ki a hibát. Ezt jelöli a piros háromszög. A zöld háromszög a költségfüggvény gradienseit számolja ki az egyes rétegekre. Az optimizáló megkapja a gradienseket és a neurális hálózat súlyait, majd javít rajtuk.

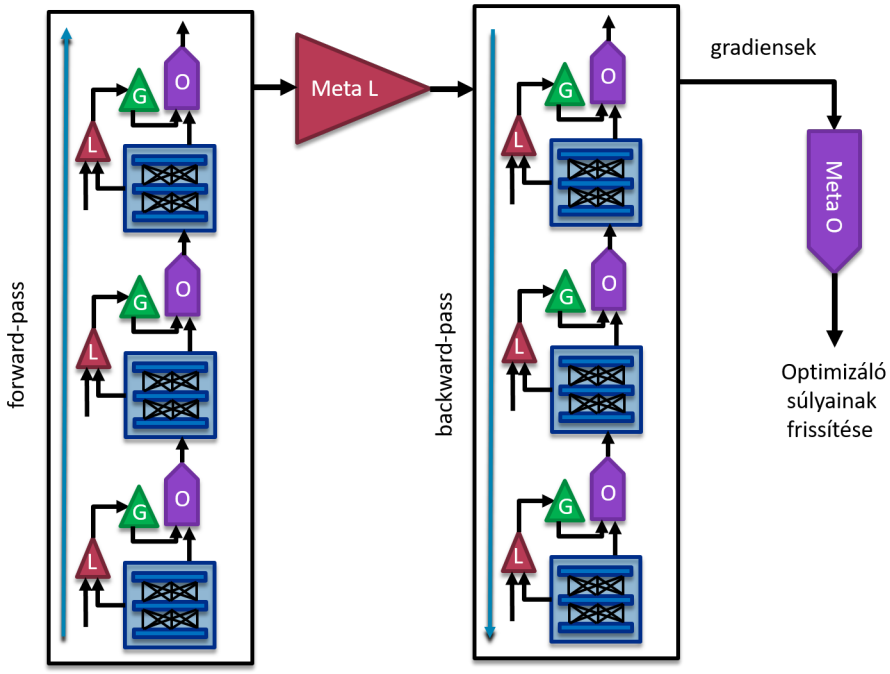


3.3. ábra. Neurális hálózat architektúra.

Tegyük fel, hogy a 3.3 ábrán látható neurális hálózatot bináris klasszifikációra használjuk és hívjuk simán modellnek. A modell kezdeti paramétereit minden iterációban az optimizáló frissíti. A metatanítás az optimizációs algoritmus szempontjából azt jelenti, hogy az optimizáló állítható paramétereit nem mi állítjuk be kézzel, hanem a feladatot átadjuk egy másik modellnek, amit megtanítunk arra, hogy ezeket optimizálja miközben az

eredeti modellünk tanul. Tehát az optimizáló az eredeti modellünk súlyait állítja, miközben a másik modellünk az optimizáló paramétereit javítja.

Ezzel absztrakciós szintet léptünk, ezt a másik modellt nevezzük metamodellnek. A metamodellnek ugyanúgy lesz meta-költségfüggvénye, meta-gradiensei és meta-optimizálója. Ugyanakkor látnunk kell, hogy ezt a metaoptimizálót is tekinthetjük modellnek és léphetünk feljebb metaszinteket, de előbb utóbb szükség lesz egy konkrét meta-optimizálóra, ami az alatta lévő optimizáló paramétereit állítja.



3.4. ábra. Optimizáló metatanítás.

A 3.4 ábra az optimizáló metatanítást szemlélteti. A konkrét modell szinten a fekete téglalapokban függőlegesen az eredeti bináris klasszifikációra használt modellünk tanítása történik, míg a téglalapok között vízszintesen a metamodell tanítása látszik.

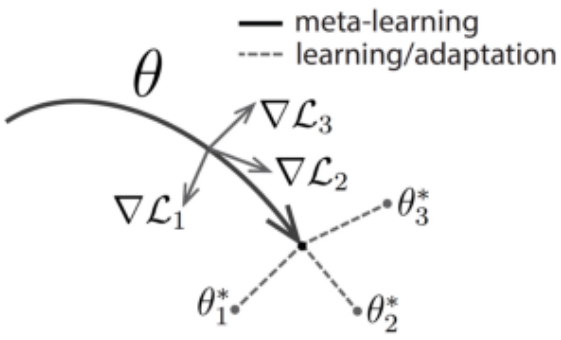
A metaköltségnek vehetjük az eredeti modell költségeinek összegét egy adott iterációig. Ez jól leírja, hogy a modell tanul-e. A metaköltség-függvény alapján kiszámoljuk a meta-gradienseket, amit átadunk a metaoptimizálónak. Ez már egy konkrét optimizáló, például ADAM. Ezután a meta-optimizáló állítja az optimizáló paramétereit úgy, hogy csökkentse a meta-költséget, vagyis javítja a tanulási folyamatot.

3.3.2. Model-Agnostic Meta Learning (MAML)

A MAML a modell kezdeti paramétereit optimalizálja úgy, hogy gyorsan tanuljon. Célja, hogy új, hasonló feladatokra kevesebb - akár egy - iterációval jó eredményt érjen el az optimizációs algoritmus. A modellt több feladathoz adaptálja a kezdeti paraméterei beállításával, így az kevés gradiens frissítés után képes megtanulni új feladatokat. Ez egy megközelítés a few-shot learning problémára.

Legyen a modell egy f_θ függvény, ahol θ jelöli a modell paramétereit. Amikor a modellt egy új $\mathcal{T}_i$ feladathoz adaptáljuk, a modell θ paramétereit változtatjuk θ_i'-re. Az adaptált paraméter egy gradiens frissítés esetén a következő:

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta) \tag{3.4}$$



3.5. ábra. A MAML algoritmus vizualizációja. (Finn et al 2017)

A 3.5 ábrán a θ jelöli a modell kezdeti súlyait. A szürke vonalak a $\nabla\mathcal{L}_1$, $\nabla\mathcal{L}_2$, $\nabla\mathcal{L}_3$ gradienseket mutatják, a θ_1^*, θ_2^*, θ_3^* pedig az adott feladathoz adaptált modell optimális paraméterei. A vastag vonal a metatanulási folyamatot mutatja. Látható, hogy a θ paraméterű modell jelenlegi helyzetében közel van mindhárom feladat optimális paramétereihez, tehát kevés gradiens frissítéssel finomhangolható bármelyikre.

Algorithm 1 Model-Agnostic Meta Learning

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β step size hyperpatameters
1: randomly initialize θ
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta\nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

Az algoritmus először veszi a $\mathcal{T}$ feladatok egy $p(\mathcal{T})$ eloszlását és a modell kezdeti paramétereit véletlen módon inicializálja. A feladatok közül kiválaszt párat és mindegyikre K tanítómintával tanítja a modellt. A tanítás során gradiens frissítésekkel kiszámolja az optimális θ_i modellparamétereket 3.4 szerint. A meta-célfüggvényt az adaptált θ_i paraméterekkel kiszámolt költségek összege adja a $\mathcal{T}_i$ feladatokon.

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}) \tag{3.5}$$

Mielőtt új $\mathcal{T}_\backslash$ feladatokat választ, sztochasztikus gradient descent módszerrel frissíti a modell kezdeti paramétereit a meta-célfüggvény szerint.

3.3.3. Reptile

A reptile algoritmus nagyon hasonlít a MAML-hoz. Ugyanúgy a hálózat kezdeti súlyait incicializálja úgy, hogy további hasonló feladatokra könnyen általánosítható legyen. A MAML-hoz képest előnye, hogy kevesebb számítást igényel, nincs szükség második deriváltak kiszámolására, csak SGD-t futtat a feladatokon.

Adott a szigma kezdeti paraméter vektor. Valamennyi iteráción keresztül választunk egy véletlen feladatot és futtatjuk rajta valahányszor az SGD algoritmust, ami a ϕ pa-

Algorithm 2 Reptile

Initialize ϕ, the vector of initial parameters
for iteration = 1, 2, ... **do**
    Sample task τ, corresponding to loss L_τ on weight vectors $\tilde{\phi}$
    Compute $\tilde{\phi} = U_\tau^k(\phi)$, denoting k steps of SGD or Adam
    Update $\phi \leftarrow \phi + \epsilon(\tilde{\phi} - \phi)$
end for

raméter vektorból a $\tilde{\phi}$-t eredményezi. Ezután javítjuk a modell kezdeti paramétereit a megadott szabály szerint.

3.4. Modell alapú metatanítás

A modell alapú metatanítás lényege az olyan modelleket tervezése, amelyek képesek kevés tanítással, gyorsan javítani a paramétereiket. Egy ismert implementáció a Memory-Augmented Neural Networks, ami neurális Turing-gépeket használ.

Köszönetnyilvánítás

Ez nem kötelező, akár törölhető is. Ha a szerző szükségét érzi, itt lehet köszönetet nyilvánítani azoknak, akik hozzájárultak munkájukkal ahhoz, hogy a hallgató a szakdolgozatban vagy diplomamunkában leírt feladatokat sikeresen elvégezze. A konzulensnek való köszönetnyilvánítás sem kötelező, a konzulensnek hivatalosan is dolga, hogy a hallgatót konzultálja.

Irodalomjegyzék

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem Villamosmérnöki és Informatikai Kar: Diplomaterv portál (2011. február 26.). http://diplomaterv.vik.bme.hu/.

[2] James C. Candy: Decimation for sigma delta modulation. 34. évf. (1986. 01) 1. sz., 72–76. p. DOI: 10.1109/TCOM.1986.1096432.

[3] Gábor Jeney: Hogyan néz ki egy igényes dokumentum? Néhány szóban az alapvető tipográfiai szabályokról, 2014. http://www.mcl.hu/~jeneyg/kinezet.pdf.

[4] Peter Kiss: Adaptive digital compensation of analog circuit imperfections for cascaded delta-sigma analog-to-digital converters, 2000. 04.

[5] Wai L. Lee – Charles G. Sodini: A topology for higher order interpolative coders. In *Proc. of the IEEE International Symposium on Circuits and Systems* (konferenciaanyag). 1987. 4-7 05., 459–462. p.

[6] Alexey Mkrtychev: Models for the logic of proofs. In Sergei Adian – Anil Nerode (szerk.): *Logical Foundations of Computer Science*. Lecture Notes in Computer Science sorozat, 1234. köt. 1997, Springer Berlin Heidelberg, 266–275. p. ISBN 978-3-540-63045-6. URL http://dx.doi.org/10.1007/3-540-63045-7_27.

[7] Richard Schreier: *The Delta-Sigma Toolbox v5.2*. Oregon State University, 2000. 01. http://www.mathworks.com/matlabcentral/fileexchange/.

[8] Ferenc Wettl – Gyula Mayer – Péter Szabó: *LaTeX kézikönyv*. 2004, Panem Könyvkiadó.