UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI

Şcoala Doctorală de Electronică, Telecomunicaţii şi Tehnologia Informaţiei

Decizie nr. 718 **din** 02-08-2021

REZUMAT TEZĂ DE DOCTORAT

Ing. Liviu-Daniel ŞTEFAN

REŢELE NEURONALE ADÂNCI PENTRU CLASIFICAREA DATELOR MULTIMEDIA

DEEP NEURAL NETWORKS FOR MULTIMEDIA CLASSIFICATION

COMISIA DE DOCTORAT

Prof. Dr. Ing. Mihai CIUC Univ. Politehnica din Bucureşti	Preşedinte
Prof. Dr. Ing. Bogdan IONESCU Univ. Politehnica din Bucureşti	Conducător de doctorat
Conf. Dr. Ing. Ruxandra ŢAPU Univ. Politehnica din Bucureşti	Referent
Prof. Dr. Ing. Henning MÜLLER Univ. de Ştiinţe Aplicate din Elveţia de Vest	Referent
SR. Dr. Ing. Adrian POPESCU CEA-LIST, Franţa	Referent

BUCUREŞTI 2021

Rezultatele prezentate în această teză au fost parțial obținute cu sprijinul Ministerului Fondurilor Europene prin Programul Operational Capital Uman, Contract nr. 51675/09.07.2019, cod SMIS 125125.

Cuprins

I Introducere şi Context Teoretic **1**

1 Introducere **1**

1.1 Prezentarea domeniului tezei de doctorat 1

1.2 Scopul tezei de doctorat 1

1.3 Conţinutul tezei de doctorat 2

2 Rețele Neuronale Artificiale Profunde **3**

2.1 Rețele Neuronale Convoluționale 3

2.1.1 Structura Rețelelor Convoluționale 3

2.1.2 VGG 3

2.1.3 GoogleNet 4

2.1.4 ResNet 4

2.2 Rețele Neuronale Recurente 4

2.2.1 Rețele Neuronale de Memorie pe Termen Lung și Scurt 4

2.3 Mecanisme de Atenție 4

2.4 Practici Bune pentru Învățarea Profundă 4

2.4.1 Augmentarea Datelor 5

2.4.2 Optimizarea Hiperparametrilor 5

2.4.3 Învățare prin Transfer 5

2.4.4 Oprimizare Fină 5

2.5 Limitările Actuale ale Învățării Profunde 5

2.6 Concluzii 6

II Contribuții Personale **7**

3 Medical Applications **7**

3.1 Învățare Profundă pentru Găsirea și Clasificarea Tipurilor de Tuberculoză pentru un Tratament Țintit 7

3.1.1 Introduction 7

3.1.2 Metoda propusă 7

3.1.3 Rezultate 8

3.1.4 Concluzii 8

3.2 Învățarea Profundă pentru Predicția de tip Multi-Ipoteză în Imagini Medicale 9

3.2.1 Introducere . . . 9
3.2.2 Metoda propusă . . . 9
3.2.3 Rezultate . . . 10
3.2.4 Concluzii . . . 10
3.3 Învățare Profundă pentru Citirea Buzelor: Către Citirea Independent de Vorbitor pentru Limba Română . . . 10
3.3.1 Introducere . . . 10
3.3.2 Metoda propusă . . . 11
3.3.3 Rezultate . . . 11
3.3.4 Concluzii . . . 12

4 Supraveghere Video **12**
4.1 Detecția și Re-Identificarea Persoanelor . . . 12
4.1.1 Introducere . . . 12
4.1.2 Metoda propusă . . . 12
4.1.3 Rezultate . . . 13
4.1.4 Concluzii . . . 13
4.2 Căutarea Persoanelor . . . 14
4.2.1 Introducere . . . 14
4.2.2 Metoda propusă . . . 14
4.2.3 Rezultate . . . 14
4.2.4 Concluzii . . . 15

5 Învățarea de Ansambluri **15**
5.1 Rețele Neuronale Adânci pentru Antrenarea Ansamblurilor . . . 15
5.1.1 Introducere . . . 15
5.1.2 Metoda propusă . . . 16
5.1.3 Predicția interesului vizual . . . 17
5.1.4 Predicția scenelor violente . . . 17
5.1.5 Predicția impactului emoțional al filmelor . . . 18
5.1.6 Detecția conceptelor . . . 18
5.1.7 Studiu de ablație . . . 19
5.1.8 Conclusions . . . 21

6 Fintech **21**
6.1 Învățarea Profundă pentru Predicția Seriilor Temporale Financiare . . . 21
6.1.1 Introducere . . . 21
6.1.2 Metoda propusă . . . 21
6.1.3 Rezultate . . . 23
6.1.4 Concluzii . . . 24

7 Baze de Date și Cadre de Evaluare publice **24**
7.1 Predicția interesului . . . 25
7.1.1 Formularea problemei . . . 25
7.1.2 Analiza metodelor . . . 25
7.1.3 Analiza generală a performanțelor metodelor . . . 25
7.1.4 Analiza performantelor rețelelor neuronale profunde de ultimă generație . . . 26
7.1.5 Arhitectura MLP propusă . . . 26
7.1.6 Evaluare . . . 27
7.1.7 Rezultate și discuții . . . 27
7.1.8 Concluzii . . . 28
7.2 Detecția violenței . . . 28
7.2.1 Descrierea bazei de date . . . 28
7.2.2 Analiza metodelor . . . 28
7.2.3 Evaluarea comparativă a metodelor de ultimă generație . . . 28
7.2.4 Concluzii . . . 29
7.3 Identificarea facială . . . 29
7.3.1 Descrierea bazei de date . . . 29
7.3.2 Rezultate și discuții . . . 29
7.3.3 Evaluarea comparativă a metodelor de ultimă generație . . . 30
7.3.4 Concluzii . . . 30

8 Concluzii generale și Perspective Viitoare de Cercetare **31**
8.1 Concluzii . . . 31
8.2 Contribuții . . . 32
8.3 Publicații . . . 34
8.4 Perspective viitoare de cercetare . . . 37

Bibliografie **38**

Partea I

Introducere şi Context Teoretic

Capitolul 1

Introducere

1.1 Prezentarea domeniului tezei de doctorat

Volumele mari de date reprezintă o provocare importantă pentru analiza, procesarea și învățarea datelor, deoarece informațiile devin disponibile în cantități copleșitoare. În acest context, imaginile, videoclipurile, sunetul, textul și alte „metadate" sunt dificil de explorat și analizat. Pentru a profita de astfel de date, abordările de învățare automată și învățare profundă, în special, sunt dezvoltate pentru a analiza fie modalitățile independente, fie întregul spectru multimedia. Mai multe dintre rezultatele studiilor consolidate ale tehnicilor de învățare profundă recente arată că sunt capabile să descopere modele complexe în seturi de date masive, datorită capacităților lor superioare de extragere de reprezentări pentru date multidimensionale, care, atunci când sunt combinate cu capacitatea lor de clasificare, depășește semnificativ descriptorii și clasificatorii tradiționali. Astfel de abordări devin din ce în ce mai fezabile, trecând de la ipoteze la sisteme și aplicații tangibile în multe domenii, cum ar fi viziunea computerizată, economia, medicina, pentru a numi doar câteva. Cu toate acestea, există încă multe domenii și probleme emergente, de nișă sau complexe, care nu beneficiază de avantajele învățării profunde. Comunitățile de cercetare explorează în continuare amploarea a ceea ce poate realiza învățarea profundă. În acest context, această teză studiază și explorează metode profunde bazate pe învățare pentru clasificarea multimedia cu aplicații în teme consacrate și emergente. Toate aceste eforturi sunt direcționate pentru a forma o bază de cunoștințe pentru practicanții de învățare automată.

1.2 Scopul tezei de doctorat

Numărul de aplicații care pot fi dezvoltate folosind învățarea automată, și în special învățarea profundă, este practic nelimitat. Cu toate acestea, multe alte aplicații depășesc considerabil capacitățile tehnicilor existente de învățare profundă. Un astfel de sector este medicina, unde mai multe metode automate de înregistrare a segmentării au fost

investigate și propuse pentru a fi aplicate în medii clinice. Chiar dacă imaginile medicale oferă o mulțime de informații, acestea sunt dificil de utilizat în practica clinică datorită imprevizibilității lor inerente. Un alt exemplu este domeniul financiar, în care datele reprezentate de prețul mărfurilor, prețul acțiunilor sau rata de schimb sunt necunoscute la $t+1$, ceea ce face extrem de dificilă antrenarea rețelelor neuronale profunde, având în vedere că acestea necesită un volum vast de date.

Pe lângă provocările menționate mai sus, există un alt factor care joacă un rol important în depășirea barierelor de performanță umane prin învățarea profundă în domenii mai specializate. Aceasta este lipsa practicienilor de învățare profundă. Un motiv este natura interdisciplinară inerent a acestui domeniu, care necesită cunoștințe puternice de matematică aplicată, programare și cunoștințe specifice domeniului pentru a înțelege, construi, rula și valida abordări bazate pe învățare profundă.

Toate aceste aspecte m-au motivat să studiez aceste domenii extrem de provocatoare și să colectez, în această teză, informații valoroase cu privire la formarea rețelelor neuronale profunde supravegheate alături de bune practici capabile să stimuleze performanțele rețelelor, lecțiile învățate și lecțiile utile pentru viitor.

1.3 Conţinutul tezei de doctorat

Această teză este structurată după cum urmează: Capitolul 2 prezintă o recenzie a literaturii abordând metodele care stau la baza implementărilor de învățare profundă, acoperind elementele de bază ale rețelelor neuronale populare, și progresează cu tehnici avansate de regularizare și optimizare pentru construirea, implementarea, configurarea, antrenarea și optimizarea modelelor de rețele neuronale. Partea II prezintă contribuțiile personale în domenii de nișă, începând cu metode legate de domeniul medical ce implică detectarea tipurilor de tuberculoză, predicția rezistenței tuberculozei la tratament, și labiolectură automată („lipreading”) în Capitolul 3, urmat de metode legate de supraveghere care implică reidentificarea și căutarea persoanelor în Capitolul 4, metode legate de învățarea automată care implică tehnici de învățare cu ansambluri și de generarea automată de modele pe baza rețelelor neuronale pentru prezicerea de diverse configurații, de la clasificare și regresie binară la clasificarea multi-etichetă, obținând rezultate “state-of-the-art” în literatura de specialitate în Capitolul 5, metode legate de domeniul economic („fintech”) care implică clasificarea tendințelor stocurilor în Capitolul 6 și o serie de baze de date care sunt în prezent referința literaturii cu privire la predicția secvențelor interesante, detectarea violenței și identificarea facială, urmate de informații, observații și recomandări privind dezvoltarea sistemelor de predicție relevante pentru fiecare dintre seturile de date în Capitolul 7. Teza se încheie cu Capitolul 8, cu constatări majore și sugestii

pentru viitorul studiu, precum și o prezentare generală a lucrărilor mele publicate și a contribuției mele la acestea.

Capitolul 2

Rețele Neuronale Artificiale Profunde

Acest capitol oferă o revizuire și un sondaj în timp util asupra arhitecturilor de învățare profundă, precum și a aplicațiilor și limitărilor acestora. Scopul său este de a oferi cititorilor un cunoștințele de bază despre diverse modele și cele mai recente progrese și realizări, pentru a ușura întelegerea conceptelor și tehnicilor propuse.

2.1 Rețele Neuronale Convoluționale

Rețelele neuronale convoluționale (CNNs) sunt reprezentate de o listă de straturi convoluționale sub forma unei stive, unde fiecare strat are o reprezentare tridimensionala (3D), și anume *lățime, înălțime, adâncime*, similar cu o imagine color RGB. Acest aranjament permite rețelelor convoluționale să recunoască secvențial forme sofisticate, de la valorile inițiale ale pixelilor până la scorurile de clasă prezise. Următoarele subsecțiuni descriu astfel de arhitecturi de învățare profundă în cadrul unei revizuiri și sondaje în timp util.

2.1.1 Structura Rețelelor Convoluționale

Convoluția este o operație liniară care măsoară cât de bine se suprapun două funcții în locații diferite. CNN-urile sunt de obicei utilizate cu imagini, tensori 3D cu doi indici de coordonate spațiale și un indice de selecție a canalului.
Subeșantionarea („pooling”) este o funcție exprimată ca o matrice de dimensiune fixă care traversează toate regiunile din intrare, realizând o subeșantionare intrării, oferind o singură valoare la fiecare locație prin diferite funcții matematice, e.g., maxim, mediere, median, etc.)
Hiperparametri sunt compuși din: (i) dimensiunile filtrului („kernel”), (ii) pasul („stride”) care indică modul de baleiere a imaginii de intrare și, (iii) bordarea („padding”) care stabilește dacă imaginea va fi bordata cu valori de 0, astfel incat sa fie parcursi toti pixelii din imagine.

2.1.2 VGG

Arhitectura VGG [32] propune nuclee de convoluție mici (3×3) cu justificarea că o stivă de două filtre de 3×3 are un câmp receptiv de 5×5 și o stivă de trei straturi convoluționale de 3×3 are un câmp receptiv cu dimensiunea de 7×7.

2.1.3 GoogleNet

GoogleNet [42] este o altă rețea neuronală convoluțională populară formată din 22 de straturi cu o arhitectură relativ complexă numită „inception” care au ca scop rezolvarea problemelor impuse de principiul profunzimii rețelelor neuronale, și anume: (i) numărul mare de parametri și (ii)) resursele de calcul.

2.1.4 ResNet

ResNet [13] aduce mai departe noi îmbunătățiri, prin premisa că rețelele mai adânci trebuie să fie cel puțin la fel de eficiente ca rețelele echivalente mai puțin adânci. Acest lucru a fost realizat învățând diferența reziduală dintre funcția de intrare și funcția de translatare.

2.2 Rețele Neuronale Recurente

Rețele neuronale recurente(RNN) [28] sunt algoritmi puternici și robuști, capabili să își amintească intrarea lor utilizând o memorie internă, lucru ce le oferă capacitatea de a fi precise în prezicerea viitoarei secvențe.

2.2.1 Rețele Neuronale de Memorie pe Termen Lung și Scurt

Rețelele de memorie pe termen lung și scurt [14] reprezintă o îmbunătățire incrementală a rețelelor recurente care permit extinderea memoriei. Această îmbunătățire ajută rețeeaua să învețe dependența pe termen lung.

2.3 Mecanisme de Atenție

Mecanismele de atenție [44] sunt în esență, funcții care translatează contextul relevant sub formă de interogări și perechi cheie-valoare la o ieșire. Presupunând că există o interogare arbitrară, mecanismele de atenție învață să participe la reprezentări intermediare prin straturi de subeșantionarea a atenției prin ponderi de atenție.

2.4 Practici Bune pentru Învățarea Profundă

Această secțiune investighează metodologii comune și bune practici pentru antrenarea arhitecturilor de învățare profundă.

2.4.1 Augmentarea Datelor

Date suplimentare pot fi obținute din eșantioanele limitate existente prin diverse transformări geometrice, transformări ale spațiului de culoare, filtre de nucleu, amestecarea imaginilor, ștergerea aleatorie [31], pentru a numi câteva.

2.4.2 Optimizarea Hiperparametrilor

Principali hiperparametri sunt: (i) **Rata de învățare** care controlează cât de repede un model ajunge la convergență, (ii) **Numărul de epoci** care controlează de câte ori se actualizează ponderile, (iii) **Straturile ascunse** care conectează straturile de intrare și ieșire ale unei rețele neuronale. Rolul lor este de a aplica transformări neliniare datelor, și (iv) **Funcțiile de activare** care introduc neliniaritatea în rețea.

2.4.3 Învățare prin Transfer

Motivația învățării prin transfer este că rețeaua va putea generaliza mai repede având la dispoziție un set redus de date dacă folosește baze de date vaste din domenii conexe.

2.4.4 Oprimizare Fină

Optimizarea fină („fine tunning") este o metodologie de învățare profundă constând în utilizarea ponderilor unei rețele neuronale existente antrenate pe o sursă de date și inițializarea unui nou model al aceleiași rețele pe datele țintă din același domeniu .

2.5 Limitările Actuale ale Învățării Profunde

Fiecare instanță din învățarea profundă este reprezentată vectorial, ceea ce înseamnă că totul poate fi privit ca un punct dintr-un spațiu geometric. Intrările modelului (care ar putea include sunet, text, videoclipuri, etc.) și țintele sunt transformate într-un spațiu vectorial de intrare inițial și, respectiv, într-un spațiu vector țintă. Fiecare strat dintr-o topologie de învățare profundă efectuează doar o singură transformare geometrică asupra datelor care trec prin el de-a lungul calculului său. Când este combinată, secvența straturilor are ca rezultat o transformare geometrică foarte complicată, care încearcă să se potrivească intrării în spațiul țintă, utilizând o serie de transformări diferite. Este posibil să se parametrizeze această transformare prin actualizarea ponderilor straturilor în mod repetat, pe baza performanței rețelei la un moment dat. Este crucial să rețineți că această transformare geometrică trebuie să fie diferențiată pentru a actualiza parametrii utilizând un algoritm de optimizare pentru a găsi un minim local. O altă limitare a modelelor de învățare profundă este că acestea nu au nicio înțelegere a datelor pe care le primesc Mai mult, când vine vorba de generalizare, modelele de învățare profundă sunt competente doar în generalizarea locală, ceea ce înseamnă că se pot adapta doar la situații noi, care sunt foarte asemănătoare cu datele anterioare.

2.6 Concluzii

Învățarea profundă supravegheată a avut un impact semnificativ în ultimul deceniu, împingând sistemele de ultimă generație în recunoașterea vorbirii, viziunea computerizată, sistemele de recomandare, înțelegerea limbajului și analiza imaginii medicale, și altele. În timp ce diferitele aspecte ale învățării profunde supravegheate sunt bine cercetate, devenirea unui practicant necesită un anumit nivel de expertiză pentru a evita supraantrenarea modelului și pentru a obține rezultatele dorite. Mai mult, antrenarea modelelor supravegheate este o sarcină laborioasă și care necesită mult timp, deoarece capacitățile predictive ale unor astfel de sisteme depind de calitatea datelor. Este bine cunoscut faptul că arhitecturile supravegheate de învățare profundă necesită un vast set de date, iar aceste date pot avea o anumită probabilitate de eroare umană, afectând calitatea predicțiilor. În ciuda acestor neajunsuri, învățarea profundă supravegheată depășește limitele în multe industrii și domenii. În ceea ce privește limitările actuale, cititorul ar trebui să observe că singura realizare concretă a învățării profunde până acum a fost capacitatea de a efectua mapări de intrare-ieșire folosind o transformare geometrică continuă în prezența unor volume enorme de date etichetate de om. Deși realizarea cu succes este transformativă pentru aproape orice domeniu, este încă departe de performanțele la nivel uman, fără a aborda alte bariere precum raționamentul și abstractizarea.

Partea II

Contribuții Personale

Capitolul 3

Medical Applications

3.1 Învățare Profundă pentru Găsirea și Clasificarea Tipurilor de Tuberculoză pentru un Tratament Țintit

3.1.1 Introduction

ImageCLEF[1] este o campanie de căutare de imaginii și de evaluare comparativă în care algoritmii sunt evaluați pe picior de egalitate. Una dintre sarcinile medicale de la ImageCLEF 2017 s-a concentrat asupra procesării datelor despre tuberculoză (TB) din imagini CT toracice (tomografie computerizată). Au fost propuse participanților două subtaskuri distincte: (1) *Multi-drug resistance* (MDR) și (ii) *Tuberculosis type* (TBT). Cititorul poate găsi aici [37] lucrarea științifică.

Contribuțiile mele la aceste sarcini sunt următoarele: i) explorarea metodelor complet automate de segmentare a plămânilor în volume CT, ii) propunerea unei modalități RGB care permite aplicarea unor bune practici pentru antrenarea metodelor de învățare profundă și iii) ajustarea fină a rețelelor neuronale din starea artei cu privire la datele medicale, obținând locul doi în competiția de predicție a tuberculozei..

3.1.2 Metoda propusă

Metoda propusă folosește arhitectura GoogleNet descrisă în [42]. Această topologie a fost aleasă datorită eficienței sale care ne permite să extindem adâncimea și lățimea rețelei, fără a crește considerabil efortul de calcul. În ceea ce privește modelarea structurii volumului, o tehnică bazată pe eșantionare structurală s-a dovedit avantajoasă alături de diverse metode de antrenare pentru a aborda provocările asociate cu un număr redus de exemple. Aceste tehnici includ pre-antrenarea între modalități și augmentarea datelor cu reprezentări îmbunătățite. De asemenea, explorăm și o tehnică automată de segmentare pulmonară pentru ambele subtaskuri. Toate aceste procesări ne permit să

[1] http://www.imageclef.org/

extragem reprezentări din fiecare imagine din volumul CT al pacienților folosind o rețea convoluțională profundă. Predicția este calculată prin medierea probabilităților furnizate printr-o clasificare SoftMax.

Multi-Drug Resistant Detection

Run 1: Rank 7: Pentru antrenarea rețelei, am folosit augmentarea datelor prin eșantionarea unei subimagini de 224×224, decupată aleatoriu din cadrul selectat. În etapa de testare, având la dispoziție un volum nou de date, extragem toate cadrele și le propagăm prin rețeaua neuronală. Etichetarea este realizată prin medierea probabilităților cadrelor din volum.

Tuberculosis Types Detection

Run 1: Rank 3: În această submisie, am folosit transformarea RGB ca intrare a rețelei optimizată fin folosind setul de date ImageNet.Pentru inferență, având la dispoziție un volum CT, extragem toate cadrele și le propagăm prin rețea. Etichetarea volumului se obține prin medierea probabilităților fiecărui cadru din volum, obținute prin funcția SoftMax.Aceasta abordare produce o precizie de 46% pe setul de validare.

Run 2: Rank 8: În această submisie, am folosit o transformare în nuanțe de gri, discretizând cadrele în intervalul de la 0 la 255 printr-o transformare liniară. Am preantrenat rețeaua folosind setul de date ImageNet și modificând ponderile straturilor de convoluție rezultate ale modelului ImageNet RGB. Pentru inferență, dat fiind un volum, extragem toate cadrele și realizăm medierea probabilităților fiecărui cadru din volum, obținute prin funcția SoftMax.Aceasta abordare produce o precizie de 41% pe setul de validare.

3.1.3 Rezultate

Pentru subtaskul MDR, abordarea propusă a obținut a 7-a poziție în funcție de acuratețe (0,5352) din 28 de submisii participante. Pentru subtaskul TBT, submisiile trimise au obținut poziția a 3-a și respectiv a 8-a dintr-un total de 23 de submisii participante, cu o valoare Kappa de 0,2329.

3.1.4 Concluzii

În acest studiu, prezentăm două tehnici de clasificare a tuberculozei complet automatizate și un model automat pentru estimarea probabilității pacienților cu tuberculoză să dezvolte o formă rezistentă la medicamente. În acest context, au fost evaluate diferite metode eficiente pentru antrenarea rețelelor convoluționale neuronale. Metodele au obținut o acuratețe de 38,7% și respectiv 51 % pe setul de date MDR atunci când se utilizează procedurile de antrenare propuse.

3.2 Învățarea Profundă pentru Predicția de tip Multi-Ipoteză în Imagini Medicale

3.2.1 Introducere

Interpretarea și rezumarea vizuală au generat o atenție tot mai sporită în comunitățile vederii artificiale și procesarii limbajului natural datorită apariției înțelegerii automate a imaginilor și a explicațiilor rezonabile. În acest context, etichetarea automată a imaginilor medicale promite să ajute medicii cu informații importante, reducând în același timp impactul general al fluxului de lucru. Datorită rolului său vital, provocarea medicală de etichetare a imaginilor investigate în competiția internațională ImageCLEF din 2017 [18] avansează metodologia de translatare a informațiilor din imagini medicale în căutarea unor metodologii îmbunătățite pentru rezumarea informațiilor vizuale. Pentru mai multe informații, cititorul este direcționat către lucrarea științifică [39].

Contribuțiile mele constau dinȘ i) propunerea unui algoritm de predicție multi-ipoteză bazat pe o rețea neuronală profundă pentru imagini medicale, ii) o metodă de codare a conceptelor bazată pe o abordare de codificare „ione-hot vector" iii) explorarea Bibliotecii Naționale de Medicină pentru construirea unei ierarhii de concepte medicale.

3.2.2 Metoda propusă

Pentru a rezolva această sarcină, ne propunem să analizăm paradigma semantică care separă conceptele textuale și imaginile asociate. Având în vedere acest lucru, prezentăm o rețea neuronală convoluțonală flexibilă care acceptă un număr arbitrar de ipoteze ca intrare și produce predicții finale multi-etichetă la ieșire. Metoda noastră profită de modelul ResNet–152 [13]. În acest context, rețeaua a fost preantrenată folosind setul de date ImageNet, ce a contribuit la inițializarea ponderilor pentru antrenarea bazei de date țintă, cu un strat "Sigmoid Cross Entropy Loss" în faza de antrenament și un strat Sigmoid în faza de testare, producând o distribuție de probabilitate între clase. Etichetele au fost codificate folosind tehnica de codare "one-hot vector". În final, lista de comcepte a fost construită pentru imaginile de testare, prin stabilirea unui prag peste scorurile de predicție.

For all trials, we present the official metric, namely F1 score. For our approach, we pretrained the network on the ImageNet dataset and finetuned it on the target dataset by taking images of size 256×256 pixels without performing any data augmentation during training. No external resources are used to augment the data set in our approach.

Evaluarea experimentală s-a realizat folosind metrica oficială, și anume metrica F1. Metoda propusă a fost preantrenată folosind setul de date ImageNet, optimizată pe setul de date țintă. Sistemul nu se folosește de augmentarea datelor în timpul antrenării.

3.2.3 Rezultate

O singură submisie a fost trimisă competiției ca parte a participării noastre. Folosind protocolul descris în secțiunea anterioară, am obținut un scor F1 de 0,089.

3.2.4 Concluzii

Această secțiune introduce un cadru pentru abordarea problemei clasificării imaginilor cu multi-etichete. Cu toate acestea, datorită numărului mare de parametri care trebuie învățați, antrenarea unei rețele CNN multietichete nu este adecvată pentru seturi de date zgomotoase, cu un număr limitat de exemple de antrenare. Prin urmare, demonstrăm modul în care un CNN antrenat cu seturi de date de imagine cu o singură etichetă, cum ar fi ImageNet, poate fi utilizat pentru problema cu etichete multiple pentru a atenua problema menționată anterior.

3.3 Învățare Profundă pentru Citirea Buzelor: Către Citirea Independent de Vorbitor pentru Limba Română

3.3.1 Introducere

Citirea buzelor („labiolectura") este procesul de interpretare a cuvintelor vorbite din informații vizuale, în special viseme care sunt reprezentate de dinamica buzelor, cu multe aplicații practice, inclusiv aparate auditive și ajutarea persoanelor cu deficiențe de auz.

În acest studiu, ne propunem să exploatăm cunoștințele lingvistice existente și să generalizăm pentrum limbi noi prin învățarea prin transfer, folosind cunoștințele dintr-un domeniu sursă pentru a le aplica într-un domeniu țintă. Intuiția noastră este că astfel de tehnici vor relaxa ipoteza referitoare la faptul că datele de antrenare trebuie să respecte anumite reguli, cum ar fi independența și distribțuia similară cu datele de testare. De asemenea, am generat un set de date multilingv la scară mică numit LRM ("Lip Reading Multilingual"), bazat pe seturi de date de citire a buzelor deja disponibile pentru a ne valida ipotezele.

Contribuția mea cuprinde: i) analizarea și implementarea bunelor practici pentru antrenarea rețelelor neuronale profunde și preprocesarea datelor într-un mod benefic pentru citirea vizuală a buzelor și ii) dezvoltarea de modele de citire a buzelor vizuale pentru echipamente eterogene, folosind tehnici de reducere a parametrilor rețelelor neuronale. pentru a dovedi fezabilitatea unor astfel de sisteme în contextul cererii masive de calcul intensiv atât pentru antrenare, cât și pentru implementarea unor astfel de sisteme. Din cunoștințele noastre, am fost primul grup de cercetare care a lansat un set

Tabel 3.7 Rezultatele obținute pentru setul de date LRM multilingv, exprimate folosind metrica de acuratețe.

Model	Invdividual subset	Acc. Top-1			
		hard		hard+easy	
		Val	Test	Val	Test
D3D	LRM	0.625	0.620	0.752	0.745
	LRW	0.835	0.830	0.836	0.847
	LRRo	0.578	**0.614**	0.885	0.892
	LRW-1000	0.463	0.418	0.535	**0.496**

de date pentru citirea buzelor și modele pentru limba română. Cititorul este redirecționat către lucrarea științifică [20] pentru mai multe detalii.

3.3.2 Metoda propusă

Metoda propusă optimizează o tehnică populară de citire a buzelor, și anume D3D (DenseNet 3D) [35] urmată de un set de bune practici pentru antrenarea modelelor de citire a buzelor pentru recunoașterea vizuală a vorbirii.

În acest context, am eliminat informațiile redundante pentru a analiza doar regiunile discriminatorii. Astfel, pentru a recunoaște caracteristicile faciale, toate imaginile din secvențele video sunt trecute mai întâi prin detectorul de față MTCNN [48] și aliniate și normalizate utilizând caracteristicile detectate. Imaginile rezultate sunt augmentate folosind transformări afine pentru a augmenta baza de date și pentru a imita variațiile din seturile de date multilingve. Aceste etape au fost stabilite experimental după evaluarea unei varietăți de configurații.

Ipotezele sunt validate folosind trei seturi de date la nivel de cuvânt disponibile public, LRW [2] pentru engleză, LRW-1000 [47] pentru mandarină și LRRo [19] pentru română. Mai mult, prin combinarea subseturilor de date din engleză, mandarină și română, am realizat un set de date multilingv. Scopul acestui set de date este de a determina dacă tehnicile existente sunt capabile să învețe regulile de compoziție pentru diferite limbi și apoi să folosească acele informații pentru a învăța să completeze spațiile libere atunci când este necesar, folosind reguli de limbaj similare. Pentru toate experimentele, prezentăm trei metrici: acuratețea top-1, acuratețea top-5 și coeficientul Kappa Cohen pentru a ține cont de varietatea datelor.

3.3.3 Rezultate

Furnizăm rezultatele studiului pentru setul de date multilingv în această secțiune.

Învățarea multilingvă

Pe subsetul "hard" al bazei LRRo, strategia de transfer multilingv a depășit strategia de preantrenare cu 14,1 și respectiv 0,2 puncte procentuale. O creștere similară de 8,4 și 18,5 puncte a fost observată în subsetul "easy+hard". De asemenea, rezultatele

experimentale indică o îmbunătățire de 1,1 puncte pentru modelul D3D pe subsetul "easy+hard" al setului de date LRW-1000.

3.3.4 Concluzii

Rezultatele arată că: i) învățarea prin transfer s-a îmbunătățit cu fiecare dintre protocoalele de învățare aplicate subseturilor individuale precum și setului de date multilingv, ii) setul de date multilingv este o opțiune de considerat pentru a atenua dezechilibrul de clasă, și iii) învățarea multilingvă este independentă de lungimea cuvintelor, făcând posibilă bazelor de date multilingve pentru antrenarea modelelor pentru limbi mai puțin reprezentate în comunitatea științifică.

Capitolul 4

Supraveghere Video

4.1 Detecția și Re-Identificarea Persoanelor

4.1.1 Introducere

Reidentificarea persoanei abordează problema căutării unor persoane specifice prin intermediul camerelor de supraveghere care nu se suprapun. Acest lucru se realizează prin estimarea asemănărilor vizuale între diferite perechi persoană detectată-țintă.

Contribuția mea la acest studiu constă în propunerea unui cadru pentru re-identificarea persoanelor care efectuează atât detecția precum și clasificarea persoanelor concomitent, spre deosebire de abordările generale ce realizează doar re-identificarea, utilizând astfel un protocol simplificat, ce presupune o aliniere perfectă a coordonatele spațiale ale persoanelor din galerie. Cititorul poate găsi mai multe detalii în lucrarea științifică [40].

4.1.2 Metoda propusă

Sistemul propus cuprinde patru etape de procesare. Etapele sunt ilustrate în Figura 4.1. În prima etapă, realizăm recunoașterea persoanelor în fluxuri video prin predicția câmpurilor vectoriale care dezvăluie direct relația dintre componentele anatomice dintr-o imagine. Modelul propus a fost antrenatfolosind etichetele din setul de date MSCOCO [26]. A doua etapă presupune segmentarea componentelor corporale. Acest lucru se realizează prin propagarea imaginilor color achiziționate în etapa anterioară, și unificarea coordonatelor spațiale prin funcția maxim pentru a realiza segmentarea persoanelor din cadru. În al treilea pas, efectuăm extragerea reprezentărilor, unde imaginile corpului sunt

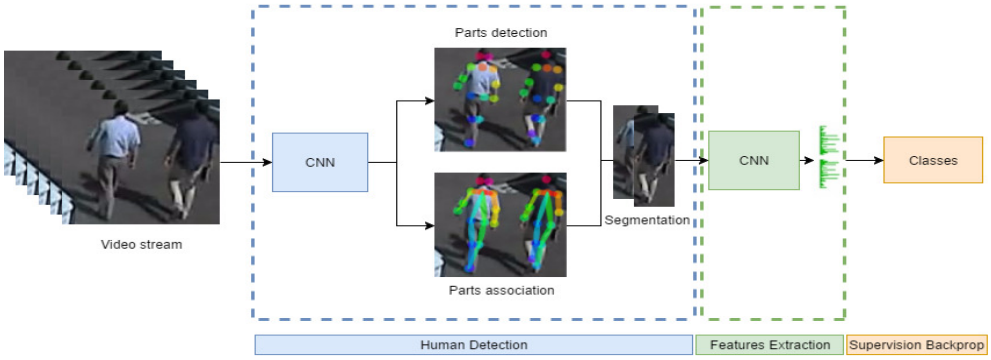


Figura 4.1 Rețeaua neuronală propusă: (1) detecție umană, (2) segmentare, (3) extracție reprezentări fiziologice, (4) clasificare SoftMax [40].

convertite în descriptori de conținut semnificativi. Acest lucru a fost realizat utilizând rețeaua VGG descrisă în [32].

În ultima etapă, folosim o clasificare SoftMax pentru a realiza antrenarea - clasificarea noilor instanțe.

Evaluare

Pentru evaluarea sistemului propus, am folosit baza de date SCOUTER împreună cu o tehnică de augmentare a datelor prin selecția a 10 imagini de dimensiuni și locații arbitrare din imaginea originală. Pentru a evalua performanța sistemului pripus, folosim o măsură globală a performanței și anume mAP.

4.1.3 Rezultate

În această secțiune, realizăm o comparație exhaustivă a sistemului propus cu sistemele "state-of-the-art" din literatură. Având în vedere specificul sarcinii, și anume supravegherea video automată, am testat două abordări: i) antrenarea de la zero, și ii) preantrenarea sistemului folosind baza de date ImageNet. Prima noastră abordare, antrenarea de la zero, atinge un scor mAP de 60,56%, în timp ce a doua abordare, preantrenarea sistemului atinge un scor mAP de 66,86%, depășind celelalte abordări existente cu 5 respectiv 11 puncte procentuale.

4.1.4 Concluzii

În această secțiune, am abordat problema reidentificării persoanelor prin dezvoltarea unei rețele neuronale profunde capabile să re-identifice cu precizie ridicată persoanele în videoclipuri cu flux multiplu dintr-o varietate de surse (interioare și exterioare).

4.2 Căutarea Persoanelor

4.2.1 Introducere

Având la dispoziție o imagine de interogare, căutarea unei persoane încearcă să localizeze persoana specificată într-o galerie de imagini utilizând imaginea de interogare. Este o generalizare a problemei de clasificare a reidentificării a persoanelor [50, 17], care se bazează pe două ipoteze și anume: i) coordonatele spațiale ale fiecărei persoane din galerie sunt furnizate și (ii) coordonatele spațiale sunt aliniate perfect. Cu toate acestea, aceste două ipoteze nu sunt valabile în realitate.

În această secțiune, profităm de progresele învățării profunde și propunem un cadru de căutare de persoane complet, care integrează arhitecturi multiple de rețele neuronale. Contribuțiile personale sunt următoarele: (i) integrarea de mecanisme de atenție în nucleul rețelei. Acest lucru permite rețelei să se concentreze pe regiuni discriminatorii, de exemplu, părți ale feței și ale corpului, sau pe diferite accesorii, cum ar fi ochelari, genți, etc., și (ii) efectuarea de transformări spațiale pentru a crește robustețea sistemului la variații spațiale. Cititorul poate găsi mai multe informații în lucrarea științifică [36].

4.2.2 Metoda propusă

În această parte, oferim o descriere globală a rețelei de căutare a persoanelor propuse. În plus, pentru a crește capacitatea predictivă a algoritmului, cercetăm, testăm și încorporăm straturi de atenție în proiectare. Metodologia pentru efectuarea unei căutări de persoane unificate constă din trei părți: (i) caracteristici globale, care sunt caracteristici de nivel scăzut extrase din întreaga imagine de intrare, (ii) propuneri de regiune pentru convertirea caracteristicilor de nivel scăzut în propuneri de persoane și (iii) caracteristici locale, care sunt caracteristici discriminatorii corespunzătoare identităților din imagini.

Astfel, am implementat și evaluat trei rețele neuronale profunde (DNN) "state-of-the-art" din literatură ca pilonuri ale arhitecturii noastre, și anume GoogleNet [42], ResNet50 [13] și DenseNet121 [16].

Evaluare

Evaluăm metoda propusă folosind trei baze de date de referință din literaturp pentru detecția și reidentificarea la scară largă a persoanelor, și anume: PRW [51], CUHK03 [25] și CUHK-SYSU [46]. Pentru fiecare set de date, am adoptat protocolul original de evaluare pe care îl oferă setul de date, și anume mAP.

4.2.3 Rezultate

Comparăm sistemul propus de căutare a persoanelor (cu sau fără utilizarea mecanismelor de atenție) folosind trei rețele profunde consacrate, și anume GoogleNet, ResNet50 și

Tabel 4.3 Eficacitatea sistemului de căutare a persoanelor exprimată în MAP (IoU $> 0,75$)

Metodă	PRW	CUHK03	CUHK-SYSU
GoogleNet	0.263	0.668	0.685
ResNet50	0.327	0.714	0.753
DenseNet121	0.335	0.703	0.778
$GoogleNet_{att}$	**0.281**	**0.694**	**0.692**
$ResNet50_{att}$	**0.347**	**0.721**	**0.775**
$DenseNet121_{att}$	**0.358**	**0.717**	**0.783**

DenseNet121 ca piloanea a sistemului pentru a evidenția utilitatea lor în căutarea de persoane. Rezultatele sunt rezumate în tabelul 4.3.

DenseNet121 depășește celelalte două rețele pe două din trei seturi de date. Tendința este valabilă pentru variațiunile cu mecanisme de atenție, care depășește în mod constant variațiunile fără mecaniseme de atenție..

4.2.4 Concluzii

Experimentele extinse arată că mecanismul de atenție îmbunătățește în mod constant performanța generală a sistemului, obținând un scor mAP de 0,358, 0,721 și 0,783 pe seturile de date PRW, CUHK03 și, respectiv, CUHK-SYSU.

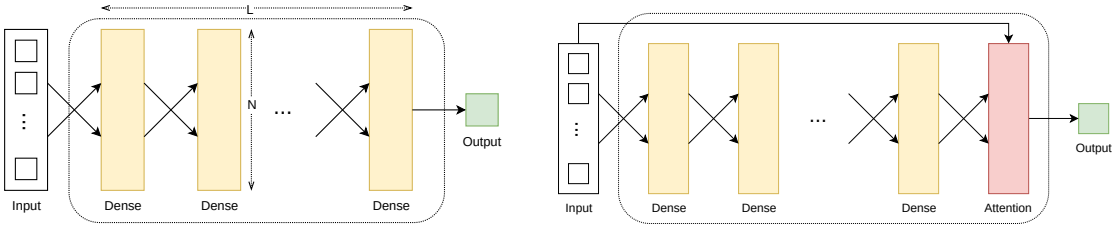
Capitolul 5

Învățarea de Ansambluri

5.1 Rețele Neuronale Adânci pentru Antrenarea Ansamblurilor

5.1.1 Introducere

Un ansamblu este definit ca o colecție de clasificatori independenți, fiecare cu propriul său domeniu de competență, și un algoritm de învățare care generează o nouă ieșire bazată pe ieșirile separate ale clasificatorilor pentru a prezice instanțele viitoare de intrare. Acest cocentp prezintă o strategie pentru fuziunea târzie bazată pe un set de sisteme. Obiectivul *ansamblurilor* este de a produce un mecanism puternic bazat pe expertiza numeroaselor clasificatoare pe care le conține. În această secțiune, prezentăm o arhitectură de ansambluri bazată pe învățarea profundă pentru a descoperi modele și corelații între răspunsurile clasificatorilor individuali.



(a) Arhitectura densă. (b) Arhitectura densă cu mecanisme de atenție.

Figura 5.2 Arhitecturi profunde de ansambluri propuse: număr variabil de straturi dense (L) și număr variabil de neuroni pe strat dens (N) [38].

Concret, contribuțiile personale constau din: i) propunerea unui nou generator de rețele neuronale profunde "end-to-end". Arhitecturile sunt construite într-o ordine progresivă, mai întâi modele simple prin variația numărului de straturi dense, numărul de neuroni pentru fiecare strat dens, și includerea sau excluderea straturilor de normalizare a mini setului de antrenare, ii) dezvoltarea de rețelele neuronale profunde adaptate pentru învățarea ansamblurilor și iii) validarea sistemelor pe o colecție largă de cadre de evaluare comune sub diferite formulări, de la clasificarea / regresia binară la clasificarea multi-etichetă. La momentul publicării, aceste metode depășesc considerabil arhitecturile de ultimă generație pe toate seturile de date evaluate. Cititorul poate accesa lucrările complete aici [4, 3, 38].

5.1.2 Metoda propusă

The proposed method follows the hypothesis that by aggregating using a deep neural network architecture, we can more efficiently describe the bias learnt by each system and the correlations across the biases, all while lowering the variance, allowing for robust retrieval. To do this, we utilize straightforward yet efficient deep neural network designs based on dense and attention ensembles.

Arhitectura densă

Pornind de la premisa că arhitecturile dense sunt aproximatori universali capabili să învețe orice funcție, ne propunem să generăm în serie variațiuni de astfel de rețele. Astfel, am definit un set de criterii pentru construirea topologiilor de rețea, care includ următoarele: i) modificarea numărului de straturi complet conectate, (ii) variația numărului de neuroni din fiecare strat complet conectat, și (iii) includerea sau eliminarea straturilor de normalizare. Diagrama arhitecturii dense implementate este prezentată în Figura 5.2a.

Arhitectura densă cu mecanisme de atenție

Pentru a crește acuratețea arhitecturii dense de ansambluri, propunem implementarea de mecanisme de atenție. În acest sens, hărțile de caracteristici sunt înmulțite cu hărți de

atenție cu valori cuprinse între 0 și 1. Diagrama implementării propuse este prezentată în Figura 5.2b.

5.1.3 Predicția interesului vizual

Setul de date MediaEval 2017 Predicting Media Interestingness [8] oferă date pentru două scenarii: (i) predicția interesului vizual in imaginii (*INT2017.Image*) și (ii) predicția interesului vizual în videoclipuri (*INT2017.Video*). Pentru antrenarea rețelelor de ansambluri, am luat în considerare toate sistemele care participă la campania de evaluare, și anume 33 de sisteme pentru INT2017.Image și 42 de sisteme pentru INT2017.Video.

Evaluare

Evaluarea se realizează utilizând următoarele scenarii de divizare a datelor de testare: (i) 75% antrenare și 25% testare (RSKF75) și (ii) 50% antrenare și 50% testare (RSKF50). Instanțele divizate sunt amestecate aleatoriu de mai multe ori pentru a obține o acoperire totală. La final sunt generate 100 de partiții. Scorurile sunt calculate prin medierea scorurilor pentru toate aceste partiții.

Rezultate

Această secțiune prezintă rezultatele celor mai performante arhitecturi. Rezultatele obținute sunt rezumate în Tabelul 5.1. Pentru datele INT2017.Image, cea mai performantă arhitectură densă folosește 10 straturi dense și 1.000 de neuroni pe strat, fără normalizarea minisetului de antrenare, obținând un mAP@10 de 0.3355 pentru RSKF75 și de 0.2316 pentru RSKF50. Adăugarea stratului de atenție a îmbunătățit și mai mult rezultatele, astfel, cea mai performantă arhitectură cu divizarea RSKF50 realizând un mAP@10 de 0,2399. Pentru datele INT2017.Video, cea mai bună performanță cu o arhitectură densă este obținută folosind 25 de straturi și 2.000 de neuroni pe strat, cu normalizarea minisetului de antrenare, obținând un MAP@10 de 0,2677 și 0,1562, pentru RSKF75 și respectiv RSKF50.

5.1.4 Predicția scenelor violente

Setul de date MediaEval 2015 Affective Impact of Movies (*VSD2015.Video*) [33] este compus din 31 de filme, 86 de videoclipuri de pe platforma YouTube și 10.900 de clipuri scurte extrase din 199 de filme de diferite genuri (până la 96 de ore)). Pentru antrenarea arhitecturilor de ansambluri, am luat în considerare toate cele 48 de sisteme antrenate în competiție.

Evaluare

Pentru evaluarea performanței, folosim metrica oficială publicată de autorii datelor (pe care au fost optimizați clasificatorii), și anume mAP. Metrica este calculată folosind instrumentul software treceval[1].

Rezultate

Rezultatele obținute sunt rezumate în Tabelul 5.1. Pentru VSD2015.Video, o arhitectură cu 5 straturi dense și 500 de neuroni pe strat a obținut cea mai ridicată performanță pentru arhitectura densă, cu un MAP de 0,6341 pe setul RSKF75, în timp ce arhitectura densă cu mecanisme de atenție a obținut un scor mAP de 0,6486.

5.1.5 Predicția impactului emoțional al filmelor

MediaEval 2018 Emotional Impact of Movies [7] este un set de date pentru recunoașterea automată a emoției în videoclipuri, în termeni de valență, excitare și frică. Pentru antrenarea arhitecturilor de ansambluri, am luat în considerare toate sistemele care au participat la campania de evaluare, și anume 30 de sisteme pentru sarcina de predicție a valenței și excitării și 18 sisteme pentru sarcina de detectare a fricii.

Evaluare

Pentru a evalua performanța modelelor propuse, am adoptat metricile oficiale propuse de autorii datelor, și anume: (i) pentru eroarea medie pătratică (MSE) și Coeficientul de corelație Pearson (PCC) pentru sarcina de predicție a valenței și excitării, MSE fiind metrica primară și Intersecția peste uniune (IoU) a intervalelor de timp, pentru sarcina de detectare a fricii.

Rezultate

Rezultatele obținute sunt rezumate în Tabelul 5.1. Pentru baza de date Arousal-Valence, cea mai performantă arhitectură DF-Dense a reușit să îmbunătățească rezultatele, ajungând la 0,0549 și respectiv 0,0626. Rețeaua DF-Attn a îmbunătățit și mai mult aceste rezultate. Pentru baza de date Fear, cea mai performantă configurație DF-Dense atinge un scor de 0,2129, crescând rezultatele "state-of-the-art" din literatură cu 35,17%. În plus, atât configurația DF-Attn crește scorul cu un IoU maxim de 0,2242, reprezentând o creștere de 42% față de performanța inițială.

5.1.6 Detecția conceptelor

ImageCLEFmed 2019 Concept Detection (*Caption*) este un set de date de etichetare automată a imaginilor și de înțelegere a scenei. Pentru antrenarea arhitecturilor de

[1] https://trec.nist.gov/trec_eval/

Tabel 5.1 Rezultate obținute pentru INT2017.Image și INT2017.Video cu metrica mAP@10, VSD2015 cu metrica mAP, Caption cu metrica F1, Arousal-Valence, cu metrica MSE și PCC și frică, cu metrica IoU .

System	Split	INT2017 Image (mAP@10)	INT2017 Video (mAP@10)	VSD2015 Video (mAP)	Caption (F1)	Arousal (MSE)	Arousal (PCC)	Valence (MSE)	Valence (F1)	Fear (IoU)
SOA	orig	0.1560	0.0930	0.3030	0.2823	0.1334	0.3358	0.0837	0.3047	0.1575
Best-LF	KF75	0.1674	0.1129	0.3920	0.2846	0.1282	0.3911	0.0769	0.3972	0.1621
DF-Dense	KF75	0.2316	0.1563	0.6192	0.3462	0.0549	0.8315	0.0626	0.8101	0.2129
	KF50	0.3355	0.2677	0.6341	**0.3740**	0.0571	0.8018	0.0640	0.7876	0.1938
DF-Attn	KF75	**0.2399**	0.1668	0.6228	0.3522	0.0548	**0.8339**	**0.0626**	**0.8107**	**0.2140**
	KF50	0.3389	**0.2750**	**0.6486**	0.3659	**0.0568**	0.8036	0.0640	0.7888	0.1913

ansambluri, am luat în considerare toate sistemele care au participat la campania de evaluare, și anume 58 de sisteme.

Evaluare

Pentru a evalua performanța modelelor propuse, am adoptat metricele oficiale propuse de autorii datelor, și anume, scorurile F1 calculate pe imagine și media calculată pentru toate imaginile. Metricile sunt calculate ca valori medii pentru toate partițiile.

Rezultate

Rezultatele obținute sunt rezumate în Tabelul 5.1. Cea mai performantă configurație DF-Dense este reprezentată de o rețea cu 5 straturi, cu 500 de neuroni pe strat și fără normalizare pe minisetup de antrenare. Această configurație atinge un scor F1 de 0,3740, crescând rezultatele "state-of-the-art" obținute în literatură cu 32,48%.

5.1.7 Studiu de ablație

Tabel 5.6 Analiza rezultatelor în corelație cu performanța și diversitatea clasificatorilor. Vă prezentăm scorurile înregistrate de cei mai performanți și cei mai puțini performanți clasificatori și respectiv, cele mai bune rezultate atât pentru RSKF75, cât și pentru RSKF50, unde ($\bar{r}$) reprezintă măsura diversității și (δ_{75}, δ_{50}). creșterea procentuală față de cel mai bun clasificator.

Run	INT2017.Image	INT2017.Video	VSD2015.Video
Best Ind.	0.1385	0.0827	0.296
Worst Ind.	0.0126	0.0396	0.0419
$\bar{r}$	0.225	0.1017	0.1997
RSKF75	0.3436	0.2799	0.6486
RSKF50	0.2399	0.1692	0.6281
δ_{75}	148.08	238.45	119.12
δ_{50}	73.21	104.59	112.19

O analiză importantă este studiul influenței diversității sistemelor asupra rezultatelor. Pentru a realiza acest lucru, folosim coeficientul de corelație Pearson. Metodologia este descrisă în versiunea extinsă a tezei. Având în vedere vectorul de intrare x_i pentru un

eșantion i de dimensiunea k, reprezentând scorurile de ieșire ale fiecărui clasificator, $\{f_0, f_1, ..., f_{k-1}\}$ (conform ecuației 5.4), decorăm fiecare element al acestui vector cu scoruri de ieșire și scoruri de corelație din cele mai similare sisteme. Prin urmare, având în vedere matricea Y (vezi ecuația 5.5) în care fiecare dintre vectorii y_i reprezintă scorurile date de clasificatorul f_i pentru toate eșantioanele, calculăm similaritatea dintre fiecare clasificator folosind coeficientul Pearson (PCC). Vectorul ordonat descendent al corelațiilor pentru fiecare vector y_i corespunde lui R_i (vezi ecuația 5.6), unde $R_{0,i}$ este coeficientul de corelație al celui mai similar sistem. $R_{1,i}$ reprezintă coeficientul de corelație al celui de-al doilea cel mai similar sistem și așa mai departe. Versiunea finală decorată a intrării este reprezentată în ecuația 5.7, unde, pentru fiecare eșantion i, perechile $(c_{0,j,i}, r_{0,j,i})$ reprezintă scorul de ieșire (c) și scorul de corelație Pearson (r) pentru cel mai similar sistem cu orice scor de sistem dat s_j. perechile $(c_{1,j,i}, s_{1,j,i})$ reprezintă al doilea sistem cel mai similar și așa mai departe. Dimensiunea vectorului decorat xd_i este de $3k \times 3$.

The decorated vector xd_i size is $3k \times 3$.

$$x_i = \begin{bmatrix} s_{0,i} & s_{1,i} & ... & s_{k-1,i} \end{bmatrix} \tag{5.4}$$

$$Y = \begin{bmatrix} y_0 & y_1 & ... & y_{k-1} \end{bmatrix} \tag{5.5}$$

$$R_i = \begin{bmatrix} R_{0,i} & R_{1,i} & ... & R_{k-2,i} \end{bmatrix} \tag{5.6}$$

$$xd_i = \begin{bmatrix} r_{3,0,i} & c_{0,0,i} & r_{0,0,i} & & r_{3,k-1,i} & c_{0,k-1,i} & r_{0,k-1,i} \\ c_{3,0,i} & s_{0,i} & c_{1,0,i} & ... & c_{3,k-1,i} & s_{k-1,i} & c_{1,k-1,i} \\ r_{2,0,i} & c_{2,0,i} & r_{1,0,i} & & r_{2,k-1,i} & c_{2,k-1,i} & r_{1,k-1,i} \end{bmatrix} \tag{5.7}$$

Prin obținerea vectorul coeficientului de corelație R_i corespunzător fiecărui clasificator f_i (vezi ecuația 5.6), putem calcula valoarea medie a acestui vector $barR_i$. În final, corelația generală pentru toate sistemele, $\bar{r}$, este calculată ca medie a tuturor valorilor $\bar{R}_i$, utilizată ca măsură a diversității pentru clasificatori. La fel ca și coeficientul de corelație Pearson, valori mai mici de $\bar{r}$ indică o corelație mică între sisteme și, prin urmare, mai multă diversitate. Tabelul 5.6 prezintă rezultatele acestei analize. Rezultatele arată că, pentru RSKF75, cea mai mare creștere (δ_{75}) este înregistrată de sistemele INT2017.Video, care măresc rezultatul celui mai bun clasificatori cu 238,75%. Acest lucru este interesant deoarece clasificatorii pentru INT2017.video sunt cei mai diversi din toate cele trei baze de date, $barr = 0,1017$. Cu toate acestea, observația nu este în concordanță cu RSKF50, unde al doilea cel mai divers clasificator pe baza VSD2015.video, $\bar{r} = 0.1997$, arată cea mai mare creștere, $\delta_{50} = 112,19\%$, în timp ce clasificatorii INT2017.Video ocupă locul doi, cu $\delta_{50} = 104,59\%$. Aceasta ar putea fi o confirmare că RSKF50 poate furniza prea puține eșantioane pentru task-uri.

5.1.8 Conclusions

În acest capitol am dezvoltat o metodă de învățare de ansambluri care utilizează arhitecturi compuse din straturi dense și de atenție. Avantajul principal al arhitecturii propuse este capacitatea de a găsi conexiuni între clasificatori în mod automat. Rezultatele au demonstrat o îmbunătățire semnificativă față de sistemele și abordările "state-of-the-art" din literatură, cu multe cazuri depășind considerabil rezultatele acestora.

Capitolul 6

Fintech

6.1 Învățarea Profundă pentru Predicția Seriilor Temporale Financiare

6.1.1 Introducere

Predicția piețelor bursiere este de mare interes în domeniul financiar, cu multe studii empirice care sugerează că piețele bursiere sunt predictibile într-o oarecare măsură. Această secțiune propune studiul potențialului învățării profunde ca mecanism de predicție a rentabilității stocului. De asemenea vizăm atât generarea de date sintetice, cât și predicția tendințelor[1]. Cititorul poate accesate aici [11, 10] lucrările științifice detaliate.

6.1.2 Metoda propusă

Fie U un univers bursier complet, ne propunem să analizăm fiecare entitate și să căutăm o funcție de predicție f pentru a prezice rentabilitatea fiecărei entitate la momentul $t+1, r_t+1.$ dată fiind reprezentarea f_t, care poate fi o transformare liniară sau neliniară a datelor originale R_t extrase la momentul t. În cercetarea noastră, folosim rentabilitatea din trecut a entităților. Considerând entitățile M din universul U și pasul de predicție g, R este dat de $R_t = \left[r_{1,t}, ...r_{1,t-g+1}, ..., r_{M,t}, ..., r_{M,t-g+1}\right]^T$.

Generarea de date sintentice

Sintetizăm serii temporale cu proprietăți cantitative comparabile cu cele din seriile temporale financiare originale, vizând proprietățile statistice, cum ar fi proprietățile de distribuție, dependență, secțiuni transversale în limite similare setului de date original. Întregul proces este descris în [11, 10]. Modelul nostru GAN sintetizează o fereastră

[1] Această lucrare este rezultatul cercetării finanțate de Hana TI.

întreagă spre deosebire de abordările clasice care sintetizează eșantionul (n + 1), pe baza a n eșantioane precedente.

Clusterizarea entităților prin metode statistice nesupervizate

Am selectat un algoritm de învățare nesupravegheat în detrimentul clasificărilor din industrie ca GICS[2], NAICS[3], Finviz[4] etc., deoarece datele sintetice (date generate prin rețele GAN) nu sunt cunoscute de metodele de clasificare menționate mai sus, prin urmare, pentru a depăși acest neajuns, este necesar să folosim un algoritm automat pentru a grupa universul real și sintetic de entități. În acest context, am aplicat algoritmul de clusterizare KMeans pentru a grupa întregul univers de entități în funcție de cât de aproape sunt rentabilitățile normalizate de media secțiunii transversale ale centroizilor, urmând procedura din [21]. Fie N numărul de observații, d zilele de tranzacționare și R_{este} randamentele zilnice ale entităților, unde $i = 1,\ldots,N$ și $s = 1,\ldots,d$, grupăm randamentele normalizate R_{is}, unde $\hat{R}_{is} = \frac{R_{is}}{\sigma_i u_i}$ și $u_i = \frac{\sigma_i}{v}$. Pentru toate $u_i < 1, u_(i) \equiv 1$ și Median $(\cdot)$ și MAD $(\cdot)$ sunt secțiuni transversale. Abaterea standard este calculată cu o fereastră în trecut de 100 de zile, iar clusterizarea se realizează cu o fereastră în trecut de 1000 de zile și cu un pas de deplasare de 30.

Etichetare statistică automată

Fie P_{is} seria temporală a prețului entităților, unde $i = 1,\ldots,N$ etichetează stocurile și $s = 1,\ldots,d$ etichetează datele de tranzacționare, o serie temporală din ziua s va fi atribuită cu o etichetă corespunzătoare, notată L_{is}, în funcție de valoarea P_{is} comparativ cu medianul clusterului de care aparține. Dacă P_{is} este mai mare sau egal cu medianul, atunci $L_{is} = 1$, otherwise $L_{is} = 0$.

Metode de extragere de reprezentări

Fie w dimensiunea ferestrei, T dimensiunea eșantionului și să presupunem că numărul de pași dintre ferestrele succesive este de o perioadă, seria temporală este partiționată în $N = T - w + 1$ subeșantioane. Pentru fiecare zi d, folosim dimensiunea ferestrei, w. Prima fereastră conține observații pentru perioada 1 până la w, a doua fereastră conține observații pentru perioada 2 până la $w + 1$ și așa mai departe. Această caracteristică ne permite să construim o linie de bază prin estimarea modelului de predicție folosind fiecare sub-eșantion de randamente pentru fiecare fereastră.

Filtrarea și redicerea dimensionalității a datelor

Datorită fluctuației pieței bursiere, precum și a și a zgomotului de tranzacționare, datele financiare au o structură complicată de iregularități. Ignorarea acestor anomalii ar putea

[2]https://www.msci.com/gics
[3]https://www.census.gov/naics/
[4]https://finviz.com/

duce la rezultate eronate de extragere a datelor și de modelare statistică. Ca rezultat, pentru a exploata reprezentări mai semnificative, ne propunem să reducem și să reducem dimensionalitatea datelor folosind o structură autoencoder SAE.Pentru această structură, am dimensiunea stratul ascuns la 16 și adâncimea SAE la 4.

Predicție

Au fost puse în aplicare și testate trei variante de rețele neuronale profunde, și anume un perceptroni multi-strat (MLP), o rețea neuronală de convoluție separabilă în serie 1D bazată pe ResNet, și o rețea bidirecțională de memorie pe termen scurt și lung (BiLSTM).

Evaluare

Simulările au fost realizate folosind date ce cuprind o perioadă de 17 ani, 7 ani de date pentru antrenare și 1 anul de testare. printr-un un protocol divizat în două părți i) antrenare și testare folosind date reale, ii) antrenare folosind date reale și sintetice și testare folosind date reale.

Tabel 6.1 Rezultatele predicției rețelei MLP pentru 50 de epoci exprimate în acurateţ̦e.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	49.38%	48.92%	50.04%	50.01%	47.64%	49.32%	50.04%	49.93%	49.41%
2	49.34%	49.61%	50.04%	50.03%	49.41%	49.65%	49.45%	49.75%	49.66%
3	49.81%	50.04%	50.03%	49.89%	49.84%	49.91%	49.58%	50.00%	49.89%
...	...	...	...	...	...	...	....	...	...
50	50.25%	50.50%	50.28%	50.63%	50.31%	50.16%	50.51%	50.78%	50.43%

Tabel 6.2 Rezultatele predicției rețelei ResNet1D pentru 50 de epoci exprimate în acurateţ̦e.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	49.21%	49.43%	50.04%	50.38%	49.88%	50.20%	50.02%	50.11%	49.91%
2	49.19%	49.21%	50.01%	50.53%	50.15%	50.12%	50.02%	50.09%	49.92%
3	49.37%	49.48%	50.06%	50.51%	50.13%	50.15%	50.01%	50.19%	49.99%
...	...	...	...	...	...	...	....	...	...
50	51.84%	50.44%	50.56%	51.34%	51.14%	51.05%	51.10%	51.93%	51.18%

Tabel 6.3 Rezultatele predicției rețelei BiLSTM pentru 50 de epoci exprimate în acurateţ̦e.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	49.91%	50.14%	50.44%	50.15%	50.01%	49.74%	50.35%	49.44%	50.02%
2	49.64%	50.12%	50.41%	50.17%	50.04%	49.11%	50.31%	49.91%	49.96 %
3	50.05%	50.14%	50.39%	50.11%	49.85%	49.89%	50.31%	50.11%	50.11 %
...	...	...	...	...	...	...	....	...	...
50	51.16%	51.55%	51.81%	51.52%	52.02%	51.38%	51.62%	52.04%	51.64%

6.1.3 Rezultate

Am folosit [23] ca model de referință din literatură. Manuscrisul testează 1.000.000 de modele de învățare automată pentru predicția direcțională folosind date din 1993 până în 2008 pentru a prezice direcția SPY ETF obținând o acurateţ̦e de 51,5%.

Tabel 6.4 Rezultatele predicției rețelei BiLSTM pentru 50 de epoci folosind la antrenare date mixte (sintetice + reale), și la testare date reale, exprimate în acuratețe.

Epoch	2012-2020	2011- 2019	2010- 2018	2009- 2017	2007- 2016	2006- 2015	2005- 2014	2004- 2013	Avg. Acc
1	50.21%	49.51%	50.57%	51.01%	50.45%	50.14%	48.53%	51.56%	50.25%
2	50.19%	50.01%	50.61%	51.13%	50.45%	50.26%	48.24%	51.43%	50.29%
3	50.20%	49.95%	50.04%	50.47%	50.17%	50.11%	51.19%	50.49%	50.33%
...	...	...	...	...	...	...	....	...	...
50	51.51%	52.82%	52.81%	52.42%	52.85%	51.86%	52.49%	52.91%	52.46%

Antrenare și testare cu date reale

Rețelele MLP, ResNet1D și BiLSTM am obținut o acuratețe generală de 50,43%, 51,64% și respectiv 52,46% după 50 de epoci.

Antrenare cu date mixte (reale+sintetice, testare cu date reale

Am reantrenat rețeaua BiLSTM folosind un mix de date reale și sintetice și testat strict pe date reale, obținând o acuratețe generală de 52,46% după 50 de epoci. Folosind datele sintetice, am crescut acuratețea de la 51,64% la 52,46%. Rezultatele sunt semnificative, raportându-ne la rezultatele de referință din literatură.

6.1.4 Concluzii

Această secțiune prezintă un cadru bazat pe rețele neuronale adânci pentru predicția seriilor temporale financiare ale entităților performante și subperformante. Trei topologii au fost dezvoltate și implementate, și anume, o rețea MLP, o variațiune 1D a popularei rețele ResNet și o rețea BiLSTM. Topologiile au fost testate în două setări, antrenare și testate pe date reale și respectiv, antrenare cu un amestec de date reale și sintetice și testate pe date reale. Rezultatele arată că datele sintetice au un impact pozitiv, îmbunătățind acuratețea generală de la 51,64% până la 52,46%.

Capitolul 7

Baze de Date și Cadre de Evaluare publice

Acest capitol prezintă contribuția mea la analiza, generarea sau lansarea publică a unor colecții de seturi de date, și anume: (i) $Interestingness10k$ [6], pentru predicția interesului în datele multimedia; (ii) $VSD96$ [5], pentru detectarea scenelor violente, și (iii) $FaVCI2D$ [27] pentru verificarea feței cu impostori dificili și date demografice diversificate. Mai mult, oferă informații, observații și recomandări cu privire la dezvoltarea sistemelor pentru a prezice reprezentările relevante pentru fiecare dintre seturile de date.

7.1 Predicția interesului

Dataset description

Setul de date Interestingness10k [6] este un set de date puglid și un cadru de test pentru prezicerea interesului imaginilor și videoclipurilor. Acest set de date a fost evaluat și validat în timpul campaniei MediaEval Predicting Media Interestingness în 2016 și 2017. Contribuțiile mele la acest set de date cuprind: i) analizarea tehnicilor utilizate și a capabilităților acestora pe setul de date *Interestingness10k*, care permit deducerea tendințelor generale în ceea ce privește cele mai performante sisteme ii), oferirea de informații despre capacitățile celor mai noi rețele neuronale profunde, prin analiza performanței arhitecturilor de ultimă generație cu privire la setul de date *Interestingness10k*, iii) furnizarea unei serii de recomandări cu privire la performanțele sistemelor "state-of-the-art", iv) analizarea și vizualizarea modului în care algoritmii interpretează interesul, v) dezvoltarea unei metode de învățare deansambluri bazate pe metodele prezentate în cadrul campaniei MediaEval Predicting Media Interestingness și, vi) participarea la procesul de etichetare a datelor.

7.1.1 Formularea problemei

Conform formulării problemei, am identificat 4 abordări principale, după cum urmează: mai mult de 52% din abordări folosesc tehnici de *clasificare*, 31% folosesc metode de tip "ranking" (este demn de remarcat faptul că aceste abordări obțin rezultate generale mai bune în toate cazuri comparativ cu clasificarea (Mann-Whitney-U $p < 0,005$)), 15% folosesc metode de regresie și 2% folosesc metode hibrid.

7.1.2 Analiza metodelor

Conform problemei de clasificare, am identificat următoarele abordări: 30% folosesc SVM, 28% Rețele neuronale profunde (DNN), 13% Ranking, 12% Regresie, 6% Hibrid, 6% Rețele neuronale (NN), 6% metode bazate pe distanță, 1% ansambluri, și 1% metode statistice. Privind performanțele lor generale, DNN-urile, NN și abordările hibride au fost cele mai performante.

7.1.3 Analiza generală a performanțelor metodelor

Pentru imagini, se remarcă abordările bazate pe DNN-uri și NN, cu scoruri mAP medii de 0,2460 și respectiv 0,2405. Pe de altă parte, pentru video, abordările hibride și abordările bazate pe SVM se remarcă ca fiind cele mai performante, cu scoruri media mAP de 0,1867 și respectiv 0,1822.

7.1.4 Analiza performantelor rețelelor neuronale profunde de ultimă generație

Pentru a cuantifica capacitățile actuale ale rețelelor neuronale profunde de ultimă generație, evaluăm performanța a trei arhitecturi recente de clasificare a imaginilor și a videoclipurilor. Am selectat pentru imagine arhitecturile ResNeXt-101-32x48d, PNASNet-5 și ResNet-50, împreună cu tehnici avansate de augmentare prezentate în [43], iar pentru datele video, am selectat rețelele InceptionV3 En3, IR-CSN-152 și R(2+1)-18 . Analiza rezultatelor arată că aceste rețele neuronale profunde nu obțin cele mai bune rezultate.

Tabel 7.3 Performanțele rețelelor neuronale artificiale profunde de ultimă generație antrenate pe baza de dateInterestingness10k.

	Metodă	2016 (mAP)	2017 (mAP@10)
Image	bestME	0.2336	0.1385
	bestSoA	0.2485	0.1560
	FixResNet50 [43]	0.1906	0.1099
	FixPNASNet-5 [43]	0.1981	0.1233
	FixResNeXt-101-32x48d [43]	0.2273	0.1410
Video	bestME	0.1815	0.0827
	bestSoA	0.1815	0.0930
	IR-CSN-152 [12]	0.1577	0.0629
	R(2+1)-18 [12]	0.1579	0.0644
	GSM-InceptionV3-En3 [41]	0.1738	0.0821

Pentru a înțelege modul în care algoritmii de învățare profundă interpretează instanțele vizuale și, astfel, cum încearcă să prezică interesul, am calculat Grad-CAM [30] și Guided Backpropagation [34]. Exemple relevante sunt prezentate în Figura 7.2. Rezultatele arată că, în multe cazuri, modelul se concentrează pe subiectul principal, dar mai ales pe elementele adiacente acestuia, arătând o înclinație pentru detectarea contextului care înconjoară subiectul principal. Acest lucru este valabil și pentru subiecții umani, deoarece analiza Grad-CAM arată activarea rețelei pe fețele umane, dar și de multe ori în jurul feței. Teorizăm că această concentrație de caracteristici utile pe fețele și în jurul acestora poate reprezenta o influență pozitivă asupra rezultatelor finale, deoarece fețele transmit emoții.

7.1.5 Arhitectura MLP propusă

Propunem o tehnică de fuziune târzie bazată pe o arhitectură MLP profundă. Metoda noastră este condusă de capacitatea stratului dens de a descoperi modele și conexiuni între ieșirile individuale ale sistemelor. Propunem să modelăm bias-uș învățată de fiecare sistem, precum și corelațiile dintre bias-urile pentru a îmbunătăți performanța generală a sistemului agregat.

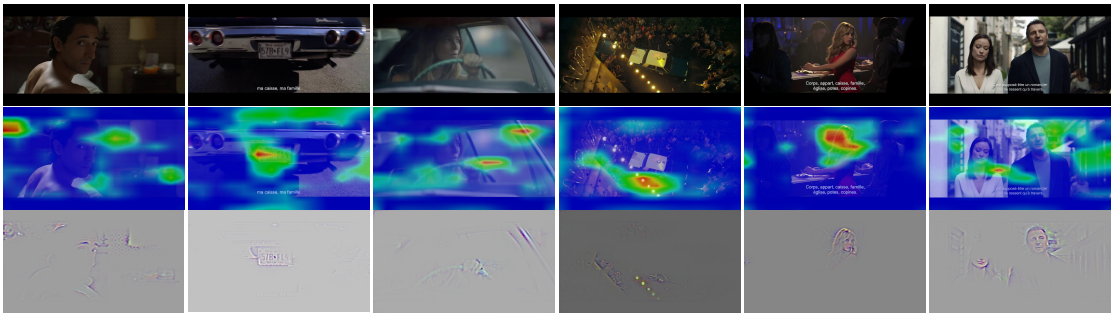


Figura 7.2 Analiza Grad-CAM a interpretării rețelei a informațiilor vizuale în cazul imaginilor prezise ca interesante. Rândul de sus prezintă instanțele originale, rândul din mijloc prezintă imaginea de ieșire Grad-CAM care descrie „regiunile discriminante de clasă", iar rândul de jos prezintă ieșirea Guided Backpropagation care descrie caracteristicile care au contribuit cel mai mult la decizia de clasă [6].

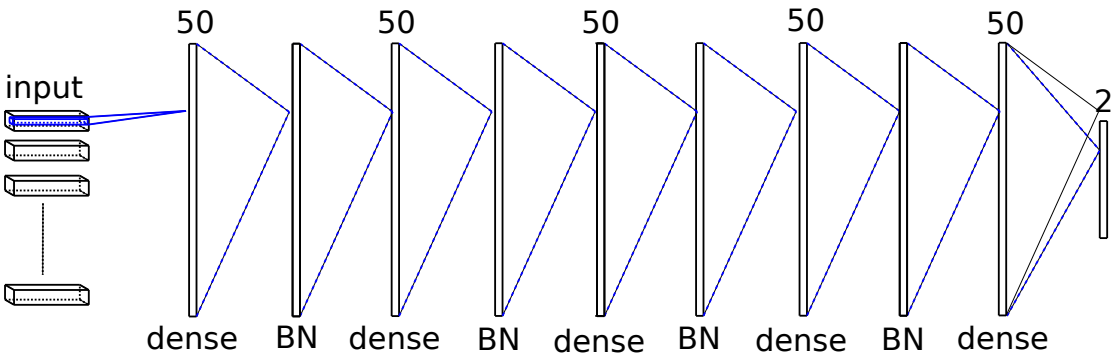


Figura 7.3 Prezentare generală a schemei de fuziune propusă [6].

După testarea cu diverse arhitecturi și parametri, am ajuns la următoarea configurație: o rețea profundă compusă din 10 straturi: 5 straturi dense activate de funcția ReLu, fiecare urmată de un strat de normalizare (bn). Rețeaua folosește scorurile de predicție de interes ale clasificatorilor ca elemente de intrare în faza de antrenament pentru a învăța decizii comune complexe. Figura 7.3 descrie topologia rețelei propuse.

7.1.6 Evaluare

Utilizăm două scenarii de partiționare a bazei de date: i) 75% antrenare și 25% testare (*RSKF75*) și (ii) 50% antrenare și 50% testare (*RSKF50*). Partițiile sunt randomizate, rezultând 100 de partiții. Valorile oficiale sunt calculate ca medii ale acestor partiții.

7.1.7 Rezultate și discuții

Rezultatele obținute cu sistemul MLP propus sunt analizate și comparate cu cele obținute de sistemele de ultimă generație din cadrul MediaEval și din literatura de specialitate. În acest context, sistemul de ansamblu depășește clasificatorii "state-of-the-art".

7.1.8 Concluzii

O constatare majoră este că, indiferent de cât de puternic este un sistem, ansamblul multor sisteme, chiar și cu performanța medie individuală, îmbunătățește performanța. Astfel, am reușit să creștem performanța aproape în fiecare caz după testarea cu numeroase abordări de fuziune, incluzând fuziunea simplă a scorurilor de sistem, tehnicile Boosting și sistemul MLP de fuziune propus. Sistemul propus îmbunătățește predicția pe imagini cu 105% față de rezultatele de top din literatură, crescând MAP de la 0,156 la 0,3202 și predicția video cu 184%, crescând MAP de la 0,093 la 0,2646.

7.2 Detecția violenței

Setul de date VSD96 [5] este un set de date public și un cadru de evaluare standardizat pentru detectarea secvențelor violente în videoclipurile Cinematic și YouTube. Acest set de date a fost validat în cadrul competițiilor MediaEval din 2011 până în 2015.

Principalele mele contribuții la acest set de date sunt după cum urmează: (i) analizarea tehnicilor de clasificare utilizate și a capacităților acestora pe setul de date *VSD* care permit deducerea tendințelor generale în ceea ce privește cele mai performante sisteme și (ii) o analiză aprofundată a metodelor reprezentative din literatură, antrenate și testate pe datele VSD96 care nu au fost trimise competiției MediaEval Impact Affective of Movies, Violent Scenes Detection.

7.2.1 Descrierea bazei de date

Setul de date VSD96 constă din filme proeminente de la Hollywood care acoperă o mare varietate de genuri, subiecte și niveluri de violență, variind de la extrem de violente până la filme fără violență.

7.2.2 Analiza metodelor

Analizăm performanța a peste 240 de sisteme implementate pentru VSD. S-au format următoarele grupuri de metode: 65% SVM, 13% metode statistice, 8% rețele neuronale (NN(, 8% metode hibride, 2%, analiză discriminantă 2% rețele neuronale profunde (DNN), 1% metode bazate pe clusterizare, și 1% metode nesupervizate. Privind performanța generală, am observat că metodele hibride au obținut cele mai bune rezultate - care ar putea fi din cauza naturii inerente multi-modale a videoclipurilor. Aceste metode sunt urmate de NN și DNN.

7.2.3 Evaluarea comparativă a metodelor de ultimă generație

Pentru a avea o analiză completă a performanței sistemelor existente, oferim o analiză aprofundată a metodelor reprezentative din literatură, instruite și testate pe datele VSD96.

Acestea nu au fost trimise la competiția MediaEval și, cel mai important, au fost dezvoltate fără constrângeri de timp. Rezultatele complete sunt prezentate în manuscrisul complet.

7.2.4 Concluzii

Această secțiune a furnizat o analiză aprofundată a componentelor esențiale ale algoritmilor VSD, prin revizuirea capacităților și a evoluției sistemelor existente, cu scopul de a oferi un ghid complet al practicantului pentru această sarcină. Am analizat 236 de sisteme care au fost trimise către MediaEval și am selectat 17 sisteme de ultimă generație din literatura de specialitate care au fost testate pe datele VSD96, care constituie o referință puternică. Am analizat fiabilitatea sistemelor, tendințele generale, metodele de predicție folosite în literatură, și posibilitatea de a agrega rezultatele sistemelor într-un super-sistem ad-hoc pentru a crește performanțele sistemelor.

7.3 Identificarea facială

Face Verification with Challenging Imposters and Diversified Demographics (*FaVCI2D*) [27] este un nou set de date de identificare a feței compus din resurse distribuibile în mod liber, conforme cu reglementările privind protecția datelor. Principalele mele contribuții la acest set de date sunt după cum urmează: (i) am oferit o analiză a arhitecturilor de ultimă generație pe seturile de date de identificare a feței existente pentru a valida rezultatele obținute cu cele raportate în literatură și faptul că modelele de extragere de trăsături ce vor fi folosite pentru validarea bazei *FaVCI2D* sunt configurate corect și comparația lor ulterioară este corectă. și ii) am furnizat trasături de ultimă generație extrase din setul de date *FaVCI2D*.

7.3.1 Descrierea bazei de date

FaVCI2D include identități din 153 de țări. Numărul total de ID-uri unice este 52.411, dintre care 12.468 sunt utilizate ca perechi autentice. Numărul total de imagini din setul de date este de 64.879.

7.3.2 Rezultate și discuții

Raportăm acuratețea modelelor de extragere de trăsături în diferite în tabelul 7.5. Similitudinea dintre ID-urile perechilor de impostori variază între 1 și aleator, care este condiția standard de verificare. Dimensiunea grupului de impostori variază între 1.000 și 52.410, numărul total de ID-uri în *FaVCI2D*. La nivel global, cea mai mare performanță o obține *insightface*, în timp ce cea mai slabă performanță o obține *facenet*.

Model	Similar = 1 Imposter pool size					Similar = 10 Imposter pool size					Similar = 100 Imposter pool size					Similar = random Imposter pool size				
	1000	5000	10000	30000	52410	1000	5000	10000	30000	52410	1000	5000	10000	30000	52410	1000	5000	10000	30000	52410
insightface	97.37 ± 0.03	96.76 ± 0.04	96.56 ± 0.03	96.07 ± 0.03	95.75 ± 0.0	98.04 ± 0.02	97.58 ± 0.05	97.34 ± 0.02	96.89 ± 0.05	96.64 ± 0.0	98.62 ± 0.02	98.22 ± 0.01	98.05 ± 0.03	97.77 ± 0.05	97.50 ± 0.0	98.81 ± 0.02	98.78 ± 0.01	98.82 ± 0.01	98.81 ± 0.03	98.82 ± 0.0
ir152	93.62 ± 0.03	92.10 ± 0.07	91.36 ± 0.05	90.08 ± 0.04	89.48 ± 0.0	95.52 ± 0.05	94.11 ± 0.08	93.43 ± 0.15	92.33 ± 0.05	91.84 ± 0.0	97.17 ± 0.02	96.13 ± 0.07	95.54 ± 0.02	94.64 ± 0.05	94.00 ± 0.0	97.65 ± 0.03	97.66 ± 0.03	97.65 ± 0.03	97.63 ± 0.04	97.64 ± 0.0
seqface	91.73 ± 0.15	89.46 ± 0.15	88.47 ± 0.16	86.65 ± 0.17	85.61 ± 0.0	94.58 ± 0.08	92.37 ± 0.11	91.31 ± 0.14	89.76 ± 0.16	88.99 ± 0.0	97.18 ± 0.06	95.55 ± 0.02	94.56 ± 0.10	93.04 ± 0.10	92.28 ± 0.0	98.04 ± 0.03	98.03 ± 0.04	98.00 ± 0.04	97.94 ± 0.02	98.06 ± 0.0
vgg	91.52 ± 0.15	89.01 ± 0.13	87.79 ± 0.08	86.00 ± 0.11	85.28 ± 0.0	94.44 ± 0.10	92.13 ± 0.07	90.89 ± 0.17	89.19 ± 0.10	88.29 ± 0.0	97.27 ± 0.09	95.33 ± 0.08	94.44 ± 0.07	92.85 ± 0.07	91.92 ± 0.0	98.37 ± 0.04	98.35 ± 0.06	98.31 ± 0.05	98.32 ± 0.06	98.42 ± 0.0
facenet	89.74 ± 0.14	86.90 ± 0.12	85.44 ± 0.13	83.48 ± 0.08	82.61 ± 0.00	93.51 ± 0.09	90.39 ± 0.04	89.13 ± 0.11	87.07 ± 0.12	86.06 ± 0.0	97.13 ± 0.05	94.79 ± 0.08	93.52 ± 0.12	91.28 ± 0.07	90.11 ± 0.00	98.37 ± 0.07	98.37 ± 0.06	98.36 ± 0.06	98.36 ± 0.06	98.39 ± 0.0

Tabel 7.5 Acuratețea de verificare cu diferite modele și configurații.

7.3.3 Evaluarea comparativă a metodelor de ultimă generație

Următoarele metode au fost folosite în experimente: *insightface* [9], bazată pe ResNet-150, antrenată pe MS-Celeb1M folosind funcția de cost ArcFace; *ir152* [49], bazată pe ResNet-152, antrenată pe MS-Celeb1M folosind funcția de cost Focal; *seqface* [15], bazată pe ResNet-27, antrenată pe MS-Celeb1M folosind funcția de cost L2-SphereFace și optimizată pe baza de date Celeb-Seq; *vgg* [1], bazată pe SE-ResNet-50 antrenată pe MS-Celeb1M și optimizată pe VGGFace2 folosind funcția de cost Softmax; *facenet* [29] bazată pe Inception ResNet, antrenată pe VGGFace2 folosind funcția de cost SoftMax. Tabelul 7.6 prezintă rezultatele obținute pe două baze de date consacrate, și anume LFW [24] și YTF [45]. Când este disponibilă, performanța modelului original este raportată în paranteză.

Rezultatele reproduse aici sunt coerente cu cele originale.Această constatare validează faptul că extractorii de caracteristici sunt configurați corect și compararea lor ulterioară este corectă.

Model	**Training Data**	**LFW**	**YTF**
insightface	MS-Celeb1M-ArcFace	99.87 (99.80+)	97.94
ir152	MS-Celeb1M	99.76 (99.80)	97.50
seqface	MS-Celeb1M + Celeb-Seq	99.80 (99.80)	98.00 (98.00)
vgg	MS-Celeb1M + VGGFace2	99.40	96.78
pfe	MS-Celeb-1M-ArcFace	99.82	97.36
dream	MS-Celeb-1M	97.71	95.01
facenet	VGGFace2	99.55	95.12

Tabel 7.6 Accuracy (%) of tested methods on LFW and YTF.

7.3.4 Concluzii

Cu recenta directivă referitoare la Regulamentul general privind protecția datelor (GDPR) din Uniunea Europeană (UE), crearea unui nou set de date de identificare a persoanelor este extrem de laborioasă. Au existat multe întrebări ridicate de comunitățile de cercetare dacă pot crea seturi de date, cum ar fi baza de date publică MegaFace [22], sau bazele de date private DeepFace sau FaceNet, deținute de Facebook și Google, respectiv, fără nicio încălcare a directivei menționate anterior. Studiul nostru arată faptul că acest lucru este posibil. În timp ce rapoartele de performanță recente dau impresia că sarcina este aproape finalizată, studiile noastre arată în continuare că seturile de date de evaluare

existente au fost construite utilizând două decizii de proiectare simplificate. O cauză identificată este că alegerea tradițională a identității folosită pentru a construi perechi de impostori este prea simplă, deoarece, de fapt, este necesară verificarea pentru a descoperi impostori greu de identificat. O a doua cauză este lipsa diversității din punctul de vedere al genului, vârstei, locației geografice, profesia, etc.

Capitolul 8

Concluzii generale și Perspective Viitoare de Cercetare

8.1 Concluzii

Acest capitol rezumă contribuțiile mele personale din timpul programului meu de cercetare doctorală în domeniul viziunii computerizate, învățării automate, medical și financiar, pentru a numi câteva, folosind metode de învățare profundă.

Partea I prezintă domeniul tezei și motivația din spatele acestei lucrări, urmată de o recenzie a literaturii care abordează metodele care stau la baza implementărilor învățării profunde.

Partea II prezintă toate contribuțiile mele personale. Capitolul 3 prezintă contribuțiile mele în domeniul medical. Secțiunile 3.1 - 3.3 descriu o serie de abordări ce implică detectarea tipurilor de tuberculoză, predicția rezistenței tuberculozei la tratament, și labiolectură automată („lipreading”). Capitolul 4 prezintă contribuțiile mele în domeniul supravegherii. Secțiunile 4.1 - 4.2 descriu metode implică reidentificarea și căutarea persoanelor în fluxuri video. Capitolul 5 prezintă contribuțiile mele în domeniul învățării automate. Secțiunea 5.1 prezintă tehnici de învățare cu ansambluri și de generarea automată de modele pe baza rețelelor neuronale pentru prezicerea de diverse configurații, de la clasificare și regresie binară la clasificarea multi-etichetă, obținând rezultate “state-of-the-art” în literatura de specialitate. Capitolul 6 prezintă contribuțiile mele în domeniul economic („fintech”). Secțiunea 6.1 prezintă metode care implică clasificarea tendințelor stocurilor, Capitolul 7 prezintă contribuția mea la lansarea unor baze de date care sunt în prezent referința literaturii cu privire la predicția secvențelor interesante, detectarea violenței și identificarea facială, alături de informații, observații și recomandări privind dezvoltarea sistemelor de predicție relevante pentru fiecare dintre seturile de date. Secțiunea 7.1 prezintă setul de date $Interestingness10k$ [6] pentru predicția interesantității în datele multimedia. Secțiunea 7.2 prezintă setul de date $VSD96$ [5] pentru detectarea scenelor violente și Secțiunea 7.3 prezintă setul de

date *FaVCI2D* [27] pentru identificarea facială cu impostori dificili și date demografice diversificate.

8.2 Contribuții

- [C3, J1] prezintă un cadru pentru generarea și predicția de serii temporale financiare. Contribuția mea principală constă în dezvoltarea algoritmului de predicție, incluzând o abordare de clusterizare pentru segmentarea universului de entități, o rețea de tip autoencoder profundă pe mai multe niveluri pentru reducerea dimensionalității și netezirea datelor, și o rețea neuronală profundă pentru prezicerea entitățile în entități performante sau neperformante.
- [J3] prezintă un studiu cuprinzător al setului de date Interestingness10k, cu scopul de a prezice interesul în imagini și videoclipuri. Acesta este cel mai cuprinzător studiu de literatură cu privire la predicția interesului în datele multimedia în momentul publicării. Contribuțiile mele cuprind analiza tehnicilor de învățare automată folosite și a capacităților acestora pe setul de date care permite deducerea tendințelor generale în ceea ce privește cele mai performante sisteme, oferirea de informații cu privire la capacitățile celor mai noi rețele neuronale profunde prin analiza performanței arhitecturi de ultimă generație pe setul de date, vizualizarea și analizarea modului în care algoritmii interpretează interesul, oferind o serie de recomandări cu privire la performanțele sistemelor, dezvoltarea unei metode de ansamblu bazate pe submisiile trimise în competiția internațională MediaEval Predicting Media Interestingness, și participarea la procesul de adnotare al setului de date.
- [C1] prezintă o nouă bază de date pentru verificarea facială cu impostori dificili și demografie diversificată numită *FaVCI2D*. Contribuțiile mele la acest set de date cuprind analiza arhitecturilor de ultimă generație pe seturile de date de verificare facială existente pentru a valida rezultatele obținute cu cele raportate în literatură și faptul că modelele de extragere de trăsături ce vor fi folosite pentru validarea bazei *FaVCI2D* sunt configurate corect și comparația lor ulterioară este corectă și furnizarea de trasături de ultimă generație extrase din setul de date *FaVCI2D*.
- [C7, C5, B1] prezintă o arhitectură nouă de ansambluri concepută pentru a descoperi tipare și corelații între deciziile clasificatorilor individuali. Principalele mele contribuții cuprind dezvoltarea a două rețele neuronale profunde adaptate pentru învățarea ansamblurilor și definirea unui set de reguli pentru a construi arhitecturi de rețele neuronale într-un mod progresiv. Validarea algoritmilor a fost realizată pe diferite task-uri de clasificare și regresie, și anume detecția excitării și valenței în fluxuri video folosind baza de date MediaEval 2018 Emotional Impact of Movies task (regresie binară), predicția interesului în îmagini și fluxuri video folosind baza de date MediaEval 2017 Predicting Media Interestingness task (regresie și clasificare), detecția violenței în fluxuri video folosind baza de date 2015 MediaE-

val Violent Scenes Detection task (clasificare binară), detecția temei folosind baza de date MediaEval 2018 Emotional Impact of Movies task (clasificare binară) și detecția conceptelor medicale folosind baza de date ImageCLEF 2019 Medical Concept Detection task (clasificare multi-clasă, multi-ipoteze), obținând rezultate "state-of-the-art" în literatura de specialitate.

- [B3,B2, B5 B4, C2, C9] prezintă o descriere a unei serii de competiții publice pentru care am fost membru al echipei organizatoare.
- [C4] propune un nou set de date pentru citirea buzelor pentru limba română. Contribuția mea cuprinde: i) analiza și implementarea bunelor practici pentru antrenarea rețelelor neuronale profunde și preprocesarea datelor pentru a exploata citirea vizuală a buzelor și ii) dezvoltarea de modele de citire a buzelor vizuale pentru dispozitive ți echipamente eterogene, folosind tehnici de tip "Graph Lowering". Am fost primul grup de cercetare care a lansat un set de date pentru citirea buzelor și modele pentru limba română.
- [J2] prezintă un studiu cuprinzător al setului de date VSD96 cu scopul de a detecta conținut video violent. Acesta este cel mai exhaustiv studiu de literatură privind predicția evenimentelor violente atât din punct de vedere subiectiv cât și obiectiv din momentul publicării, abordând și analizând peste 250 de tehnici de învățare automată. Principalele mele contribuții la acest set de date cuprind analiza tehnicilor de clasificare utilizate și a capacităților acestora pe setul de date VSD care permit deducerea tendințelor generale în ceea ce privește cele mai performante sisteme și o analiză aprofundată a metodelor reprezentative din literatură, antrenate și testate pe datele VSD96 care nu au fost trimise campaniei de evaluare MediaEval Impact Affective of Movies - Violent Scenes Detection cu scopul de a evalua progresele domeniului și de a oferi lecții utile pentru viitor.
- In [C6], am propus o arhitectură unificată bazată pe rețele neuronale profunde pentru căutarea de persoane, cu mecanisme de atenție. La momentul publicării, tehnicile de atenție nu au fost implementate în etapa de detecție, fiind astfel o noutate în literatură la momentul publicării.
- In [C11], am propus o rețea neuronală profundă pentru detecția și reidentificare persoanelor în fluxuri video. La momentul publicării, rezultatele au depășit metodele de ultimă generație din literatură testate pe setul de date.
- In [C13], mm propus o structură CNN flexibilă, care acceptă un număr arbitrar de ipoteze la intrare pentru clasificarea imaginilor multi-etichetă ("captioning"). Rezultatele sunt validate în cadrul competiției internaționale ImageCLEF 2017Caption.
- In [C12], am propus două tehnici de clasificare a tuberculozei complet automatizate și un clasificator automat pentru estimarea probabilității pacienților cu tuberculoză să dezvolte o formă rezitentă la medicamente, toate bazate pe rețele neuronale profunde. Mai mult, am propus o abordare de transformare a imaginilor 3D

CT care permite aplicarea învățării prin transfer din alte domenii utilizând rețele neuronale profunde. Rezultatele sunt validate în competiția internațională de tuberculoză ImageCLEF 2017.

8.3 Publicații

Capitole de carte:

B1: M.G. Constantin, **L.-D. Ştefan**, B. Ionescu : Exploring Deep Fusion Ensembling for Automatic Visual Interestingness Prediction. In book Human Perception of Visual Information - Psychological and Computational Perspectives, Springer International Publishing, Eds. B. Ionescu, W. Bainbridge, N. Murray. DOI: https://doi.org/10.1007/978-3-030-81465-6, September 10, 2021.

B2: B. Ionescu, H. Müller, R. Péteri, ... , **L,-D. Ştefan**, ... , J. Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications, in Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), Bucharest, Romania, Springer Lecture Notes in Computer Science LNCS, September 21-24, 2021.

B3: B. Ionescu, H. Müller, R. Péteri, ..., **L.-D. Ştefan**, ..., A. Popescu: The 2021 ImageCLEF Benchmark: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications. In: Springer Lecture Notes in Computer Science, vol 12657, DOI: https://doi.org/10.1007/978-3-030-72240-1_72, ECIR 2021 Proceedings, March 30, 2021.

B4: B. Ionescu, H. Müller, R. Péteri,..., **L.D. Ştefan**, M.G. Constantin, "Overview of the ImageCLEF 2020: Multimedia Retrieval in Medical, Lifelogging, Nature, and Internet Applications", in Springer Lecture Notes in Computer Science, 12260, CLEF 2020 Proceedings, September 22-25, Thessaloniki, Greece, 2020.

B5: B. Ionescu, H. Müller, R. Péteri, ..., M. Dogariu, **L.-D. Ştefan**, M.G. Constantin: ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications. In Springer Lecture Notes in Computer Science, vol 12036, pp. 533-541, DOI: https://doi.org/10.1007/978-3-030-58219-7_22, ECIR 2020 Proceedings, April 14-17, Lisbon, Portugal, 2020.

Reviste:

J1: *Trimis spre recenzie:* M. Dogariu, **L.-D. Ştefan**, B.-A. Boteanu, C. Lamba, B. Kim, B. Ionescu, Realistic financial time-series generation. In Transactions on Multimedia Computing Communications and Applications (TOMM), 2021, (Q1 journal article, **Impact Factor: 3.1**).

J2: M.G. Constantin, **L.-D. Ştefan**, B. Ionescu, C.-H. Demarty, M. Sjöberg, M. Schedl, G. Gravier: Affect in Multimedia: Benchmarking Violent Scenes Detection. IEEE

Transactions on Affective Computing, DOI: https://doi.org/10.1109/TAFFC.2020.2986969, April 2020. (Q1 journal article, **Impact Factor: 10.5**).

J3: M.G. Constantin, **L.-D. Ştefan**, B. Ionescu, N.Q.K. Duong, C.-H. Demarty, M. Sjöberg : Visual Interestingness Prediction: A Benchmark Framework and Literature Review. International Journal of Computer Vision, DOI: https://doi.org/10.1007/s11263-021-01443-1, 2021, WOS:000620409500001 (Q1 journal article, **Impact Factor: 7.41**).

Conferințe:

C1: *Trimis spre recenzie:* A. Popescu, **L.-D. Ştefan**, J. Deshayes-Chossart, and B. Ionescu. Faceverification with challenging imposters and diversified demographics. In IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2022.

C2: R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, **L.-D. Ştefan**, M. G. Constantin and B. Ionescu. Overview of the 2021 ImageCLEFdrawnUI Task: Detection and Recognition of Hand Drawn and Digital Website UIs, Conference and Labs of the Evaluation Forum (CLEF), 2021.

C3: M. Dogariu, **L.-D. Ştefan**, B.-A. Boteanu, C. Lamba, B. Ionescu, Realistic financial time series data generation via generative adversarial learning. In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6-11 June, Toronto, Ontario, Canada, 2021, *ISI indexed conference*.

C4: A.-C. Jitaru, **L.-D. Ştefan**, B. Ionescu, Toward Language-independent Lip Reading: A Transfer Learning Approach. In International Symposium on Signals, Circuits and Systems (ISSCS), (pp. 1-4). IEEE, 2021, *ISI indexed conference*.

C5: M.G. Constantin, **L.-D. Ştefan**, and B. Ionescu. DeepFusion: Deep Ensembles for Domain Independent System Fusion. In International Conference on Multimedia Modeling, pp. 240-252. Springer, Cham, 2021, *ISI indexed conference*.

C6: **L.-D. Ştefan**, Ş. Abdulamit, M. Dogariu, M.-G. Constantin, and B. Ionescu. Deep learning-based person search with visual attention embedding. In 2020 13th International Conference on Communications (COMM), pp. 303-308. IEEE, 2020, WOS:000612723900053, *ISI indexed conference*.

C7: **L.-D. Ştefan**, M. G. Constantin, and B. Ionescu. System Fusion with Deep Ensembles. In Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 256-260. 2020, *ACM indexed conference*.

C8: M. Dogariu, **L.-D. Ştefan**, M. G. Constantin, and B. Ionescu. Human-Object Interaction: Application to Abandoned Luggage Detection in Video Surveillance Scenarios. In 2020 13th International Conference on Communications (COMM), pp. 157-160. IEEE, 2020, WOS:000612723900028, *ISI indexed conference*.

C9: D. Fichou, R. Berari, P. Brie, M. Dogariu, **L.-D. Ştefan**, M.G. Constantin, and B. Ionescu. Overview of ImageCLEFdrawnUI 2020: the detection and recognition of

hand drawn website UIs task. In CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS. org, Thessaloniki, Greece (September 22-25 2020). 2020.

C10: C.-A. Mitrea, M.G. Constantin, **L.-D Ştefan**, M. Ghenescu, and B. Ionescu. Little-big deep neural networks for embedded video surveillance. In 2018 International Conference on Communications (COMM), pp. 493-496. IEEE, 2018, WOS:000449526000092, *ISI indexed conference*.

C11: **L.-D. Ştefan**. I. Mironică, C.-A. Mitrea. and B.Ionescu. End to end very deep person re-identification. In 2017 International Symposium on Signals, Circuits and Systems (ISSCS) (pp. 1-4). IEEE, 2017, WOS:000425211500061, *ISI indexed conference*.

C12: **L.-D. Ştefan**, Y. D. Cid, O. A. Jiménez del Toro, B. Ionescu, and H. Müller. Finding and Classifying Tuberculosis Types for a Targeted Treatment: MedGIFT-UPB Participation in the ImageCLEF 2017 Tuberculosis Task. In CLEF (Working Notes). 2017.

C13: **L.-D. Ştefan**, B. Ionescu, and Henning Müller. Generating captions for medical images with a deep learning multi-hypothesis approach: MedGIFT–UPB Participation in the ImageCLEF 2017 Caption Task.

C14: A. Toma, **L.-D. Ştefan**, and B. Ionescu. UPB HES SO@ PlantCLEF 2017: Automatic Plant Image Identification using Transfer Learning via Convolutional Neural Networks. In CLEF (Working Notes). 2017.

Proiecte de cercetare:

R1: 2020–2024: **researcher**, project H2020 AI4Media. "A European Excellence Centre for Media, Society and Democracy", owner CERTH, Greece, partner Polytechnic University of Bucharest, axis H2020 ICT-48-2020 / Towards a vibrant European network of AI excellence centres (budget 12M Eur).

R2: 2020–2022: **researcher**, project SMARTRetail. “Enhancing and Improving Customer Experience and Services in Supermarkets via SMART Artificial Intelligence Powered Systems”, owner Softrust Vision Analytics, partner Polytechnic University of Bucharest, funded by UEFISCDI, Industry Transfer Axis, grant PN-III-P2-2.1-PTE-2019-0055 (budget 340k Eur).

R3: 2020–2022: **researcher**, project GRAVI. “Virtual Guardian: Artificial Intelligence Powered Multi-Sensor System for Automatic Securing of Areas of Interest”, owner Softrust Vision Analytics, partner Polytechnic University of Bucharest, funded by UEFISCDI, Industry Transfer Axis, grant PN-III-P2-2.1-PTE-2019-0570 (budget 350k Eur).

R4: 2020 March–2020 October: **researcher**, project Keysight 1. “Machine Learning Techniques for Generating Network Traffic Data”, owner Polytechnic University

of Bucharest, CAMPUS Research Institute, beneficiary Keysight Technologies Romania (budget 59k Eur).

R5: 2020 April–2020 July: **researcher**, project Hana 2. “Financial Data Augmentation and Forecasting Using Advanced AI Techniques”, owner “Polytechnic” Research, Development and Innovation Institute, beneficiary Hana Institute of Technology, Republic of Korea (budget private).

R6: 2019 May–2019 December: **researcher**, project NXP 1. “RISC V-based Hardware-Software System for Machine Learning Applications”, owner Polytechnic University of Bucharest, CAMPUS Research Institute, beneficiary NXP Semiconductors Romania (budget 136k Eur).

R7: 2017–2020: **researcher**, project SPIA-VA. “Technologies and Innovative Video Systems for Person Re-Identification and Analysis of Dissimulated Behavior”, owner Polytechnic University of Bucharest, partners UTI Grup, Romanian Ministry of National Defence — Military Equipment and Technologies Research Agency, public beneficiary Protection and Guard Service, funded by UEFISCDI, Solutions Axis, grant 2SOL/2017 (budget 2.2M Eur).

R8: 2017–2018: **researcher**, project SPOTTER. “Real-time IP Camera-based Intelligent Video Surveillance Security System with DROP Retrieval”, owner Polytechnic University of Bucharest, partner UTI Grup, funded by UEFISCDI, PED Axis, grant 30PED/2017 (budget 140k Eur).

R9: 2016–2018: **researcher**, project Erasmus+ CBHE, UMETECH. “University & Media Technology for Cultural Heritage”, owner University of Florence, Italy, partner Polytechnic University of Bucharest (budget 900k Eur).

8.4 Perspective viitoare de cercetare

În acest moment, majoritatea practicanților de învățare profundă folosesc manipularea datelor cu scripturi Python și apoi modificarea meticuloasă a topologiei și a hiperparametrilor pentru a obține un model funcțional. Un cercetător aventuros poate recurge la un model de ultimă generație pre-antrenat de comunitatea științifică. Din păcate, aceasta nu este configurația ideală. Învățarea profundă supravegheată poate ajuta în acest domeniu. Soluțiile care gestionează majoritatea modificărilor parametrice ale modelului pot deveni un instrument puternic, așa cum am demonstrat în experimentele noastre de învățare de ansambluri. O perspectivă mai ambițioasă încearcă să descopere o arhitectură adecvată de la zero, prin învățare prin întărire ("reinforcement") sau algoritmi genetici. O altă abordare practică pentru automatizare este de a învăța arhitectura modelului împreună cu ponderile modelului.

Bibliografie

[1] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE.

[2] Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *ACCV*.

[3] Constantin, M. G., Ştefan, L.-D., and Ionescu, B. (2021a). Deepfusion: Deep ensembles for domain independent system fusion. In *International Conference on Multimedia Modeling*, pages 240–252. Springer.

[4] Constantin, M. G., Stefan, L.-D., and Ionescu, B. (2021b). Exploring Deep Fusion Ensembling for Automatic Visual Interestingness Prediction. In *Human Perception of Visual Information: Psychological and Computational Perspectives*. Springer.

[5] Constantin, M. G., Stefan, L. D., Ionescu, B., Demarty, C.-H., Sjoberg, M., Schedl, M., and Gravier, G. (2020). Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*.

[6] Constantin, M. G., Ştefan, L.-D., Ionescu, B., Duong, N. Q., Demarty, C.-H., and Sjöberg, M. (2021c). Visual interestingness prediction: A benchmark framework and literature review. *International Journal of Computer Vision*, pages 1–25.

[7] Dellandréa, E., Huigsloot, M., Chen, L., Baveye, Y., Xiao, Z., and Sjöberg, M. (2018). The mediaeval 2018 emotional impact of movies task. *In Proc. of MediaEval 2018 Workshop*.

[8] Demarty, C.-H., Sjöberg, M., Ionescu, B., Do, T.-T., Gygli, M., and Duong, N. (2017). Mediaeval 2017 predicting media interestingness task. In *MediaEval workshop*.

[9] Deng, J., Guo, J., Niannan, X., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.

[10] Dogariu, M., Ştefan, L.-D., Boteanu, B.-A., Lamba, C., and Ionescu, B. (2021). Towards realistic financial time series generation via generative adversarial learning. In *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE.

[11] Dogariu, M., Ştefan, L.-D., Boteanu, B.-A., Lamba, C., Kim, B., and Ionescu, B. (2022). Realistic financial time-series generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*. [under review].

[12] Ghadiyaram, D., Tran, D., and Mahajan, D. (2019). Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055.

[13] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[14] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[15] Hu, W., Huang, Y., Zhang, F., Li, R., Li, W., and Yuan, G. (2018). Seqface: make full use of sequence information for face recognition. *arXiv preprint arXiv:1803.06524*.

[16] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[17] Ionescu, B., Ghenescu, M., Răstoceanu, F., Roman, R., and Buric, M. (2020). Artificial intelligence fights crime and terrorism at a new level. *IEEE MultiMedia*, 27(2):55–61.

[18] Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.-T., Cid, Y. D., Eickhoff, C., de Herrera, A. G. S., Gurrin, C., et al. (2017). Overview of imageclef 2017: Information extraction from images. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 315–337. Springer.

[19] Jitaru, A. C., Abdulamit, Ş., and Ionescu, B. (2020). Lrro: a lip reading data set for the under-resourced romanian language. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 267–272.

[20] Jitaru, A.-C., Ştefan, L.-D., and Ionescu, B. (2021). Toward language-independent lip reading: A transfer learning approach. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE.

[21] Kakushadze, Z. and Yu, W. (2016). Statistical industry classification. *Journal of Risk & Control*, 3(1):17–65.

[22] Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., and Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882.

[23] Kinlay, J. (2011). Can machine learning techniques be used to predict market direction? The 1,000,000 model test.

[24] Learned-Miller, G. B. H. E. (2014). Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst.

[25] Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159.

[26] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

[27] Popescu, A., Ştefan, L.-D., Deshayes-Chossart, J., and Ionescu, B. (2022). Face verification with challenging imposters and diversified demographics. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE. [under review].

[28] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

[29] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[30] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

[31] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

[32] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[33] Sjöberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., Schedl, M., Demarty, C.-H., and Chen, L. (2015). The mediaeval 2015 affective impact of movies task. In *MediaEval*.

[34] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

[35] Stafylakis, T. and Tzimiropoulos, G. (2017). Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*.

[36] Ştefan, L.-D., Abdulamit, Ş., Dogariu, M., Constantin, M. G., and Ionescu, B. (2020a). Deep learning-based person search with visual attention embedding. In *2020 13th International Conference on Communications (COMM)*, pages 303–308. IEEE.

[37] Stefan, L.-D., Cid, Y. D., del Toro, O. A. J., Ionescu, B., and Müller, H. (2017a). Finding and classifying tuberculosis types for a targeted treatment: Medgift-upb participation in the imageclef 2017 tuberculosis task. In *CLEF (Working Notes)*.

[38] Ştefan, L.-D., Constantin, M. G., and Ionescu, B. (2020b). System fusion with deep ensembles. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 256–260.

[39] Stefan, L.-D., Ionescu, B., and Müller, H. (2017b). Generating captions for medical images with a deep learning multi-hypothesis approach: Medgift–upb participation in the imageclef 2017 caption task.

[40] Ştefan, L.-D., Mironică, I., Mitrea, C. A., and Ionescu, B. (2017). End to end very deep person re-identification. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE.

[41] Sudhakaran, S., Escalera, S., and Lanz, O. (2020). Gate-shift networks for video action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[42] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[43] Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. (2019). Fixing the train-test resolution discrepancy. In *Advances in Neural Information Processing Systems*, pages 8250–8260.

[44] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[45] Wolf, L., Hassner, T., and Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE.

[46] Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2017). Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424.

[47] Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., Long, K., Shan, S., and Chen, X. (2018). Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8.

[48] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

[49] Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J., et al. (2018). Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216.

[50] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

[51] Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., and Tian, Q. (2017). Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376.