UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI

Şcoala Doctorală de Electronică, Telecomunicaţii şi Tehnologia Informaţiei

Decizie nr. 846 **din** 09-06-2022

REZUMAT TEZĂ DE DOCTORAT

Ing. Mihai-Sorin BADEA

DEZVOLTAREA SISTEMELOR INTELIGENTE DE INTERFAŢARE VIZUALĂ OM-MAŞINĂ

THE DEVELOPMENT OF INTELLIGENT SYSTEMS FOR VISUAL HUMAN-MACHINE INTERFACES

COMISIA DE DOCTORAT

Prof. Dr. Ing. Bogdan IONESCU Univ. Politehnica din Bucureşti	Preşedinte
Prof. Dr. Ing. Constantin VERTAN Univ. Politehnica din Bucureşti	Conducător de doctorat
Prof. Dr. Ing. Cătălin-Daniel CĂLEANU Univ. Politehnica din Timişoara	Referent
Prof. Dr. Ing. Romulus-Mircea TEREBEŞ Univ. Tehnică din Cluj-Napoca	Referent
Prof. Dr. Ing. Corneliu Nicolae FLOREA Univ. Politehnica din Bucureşti	Referent

BUCUREŞTI 2022

Cuprins

Capitolul 1

Introducere

1.1 Motivaţie

Ultimele decenii au adus o dezvoltare spectaculoasă a puterii de calcul regăsite în computerele personale şi a interconectării acestora prin intermediul Internetului. Aceste aspecte au condus la integrarea treptată şi constantă a aplicaţiilor software în majoritatea aspectelor referitoare la viaţa de zi cu zi. În acelaşi timp, domeniul inteligenţei artificiale a revenit în atenţia publicului larg cu o serie de îmbunătăţiri în zona învăţării automate. O cauză importantă a acestei reveniri este reprezentată de faptul că majoritatea datelor şi a informaţiilor sunt stocate în format digital. Deşi performanţele impresionante ale algoritmilor nu sunt explicate în totalitate de modelele teoretice existente, aplicabilitatea lor nu poate fi trecută cu vederea, iar studiul comportamentului diverşilor algoritmi utilizaţi devine critic.

Conceptele specifice inteligenţei artificiale se suprapun cu idei din alte domenii, iar această legătură este din ce în ce mai clară odată cu dezvoltarea aplicaţiilor de performanţă înaltă. Soluţiile de vedere computerizată au fost influenţate vizibil de învăţarea automată, introducerea Reţelelor Neuronale Convoluţionale reprezentând un moment marcant în istoria recentă. Ţinând cont şi de natura experimentelor care vor fi prezentate în lucrare, noţiunile teoretice vor fi frecvent tratate din perspectiva imaginilor atunci când este posibil.

Frecvent, algoritmii de învăţare automată au de îndeplinit sarcini care nu necesită mari eforturi pentru oameni. Cu toate acestea, există o multitudine de cazuri în care şi pentru o persoană ar fi necesară o instruire anterioară. Spre exemplu, aplicaţiile din zona medicală se înscriu în această categorie. De asemenea, manipularea avansată a conţinutului unei imagini nu este facilă nici atunci când sunt puse la dispoziţie unelte software specializate. Mai mult, dacă însemnătatea datelor nu este clară sau numărul de variabile aparent necorelate este mare, algoritmii automaţi pot întrece rapid performanţa umană.

1.2 Obiective

Pe baza rezultatelor foarte bune, reţelele neuronale convoluţionale au ajuns să fie o alegere populară în sarcinile de vedere computerizată. Antrenarea lor însă este un proces dificil şi mare consumator de resurse de calcul, cu o atenţie deosebită acordată cantităţii şi calităţii datelor utilizate. Din acest motiv o parte din abordările care vor fi prezentate în cadrul lucrării vor viza modul de folosire a datelor de către algoritm. De asemenea, vor fi propuse o serie de costuri suplimentare pentru a îmbunătăţi performanţa arhitecturilor neuronale.

Ca domenii principale de aplicaţii, vor fi studiate analiza automată a tablourilor, precum şi a expresiilor faciale. Deşi unele soluţii se vor baza pe anumite particularităţi ale domeniilor, vor fi prezentate şi implementări care pot fi utilizate în probleme generale de învăţare automată.

1.3 Structura lucrării

Următorul capitol se va axa mai mult pe noţiunile teoretice necesare în părtile ulterioare. După ce vor fi prezentate elemente de bază ale învăţării automate, se va acorda o atenţie deosebită reţelelor neuronale convoluţionale. Având în vedere importanţa învăţării semi-supervizate în porţiune experimentală, un capitol va fi dedicat acestora.

Prima serie de aplicaţii este axată pe analiza tablourilor, în Capitolul 5. În continuare, vor fi prezentate experimentele care au avut ca subiect central analiza expresiilor faciale. Capitolul 7 conţine concluziile eforturilor de cercetare şi încheie lucrarea.

Capitolul 2

Elemente de învăţare automată

Domeniul învăţării automate (*Machine Learning* - ML) se află într-o continuă evoluţie, ultimele două decenii fiind marcate de îmbunătăţiri semnificative în ceea ce priveşte calitatea rezultatelor. Disciplina este preocupată cu studierea algoritmilor care au drept scop rezolvarea unei sarcini bine definite, fără a exista o programare explicită a condiţiilor de rezolvare.

De-a lungul timpului au fost introduse şi modificate numeroase abordări. O parte dintre cele mai importante, relevante tematicilor prezente în lucrare vor fi prezentate în continuare, alături de alte noţiuni teoretice generale. De asemenea, vor fi detaliate şi informaţii referitoare la o serie de arhitecturi şi algoritmi notabili.

2.1 Tipuri de învăţare

Din punct de vedere al tipului de învăţare utilizat, se disting mai multe categorii de algoritmi ML, cele mai importante fiind: învăţarea supervizată, învăţare nesupervizată şi *Reinforcement Learning*. În primul caz, eşantioanele utilizate X au asociate una sau mai multe etichete y care reprezintă răspunsul dorit din partea sistemului. Cazul nesupervizat presupune că nu există etichetele y, în timp ce RL se distinge prin necesitatea componentei temporale, algoritmul aflându-se într-o interacţiune continuă cu mediul în care se află.

2.2 Descriptori de imagini

Eşantioanele utilizate de către algoritmi sunt reprezentate într-o formă numerică, în general ca un vector de *trăsături*. În cazul imaginilor, eşantioanele sunt descrise de valorile pixelilor din componenţa lor. În majoritatea algoritmilor, aceste valori sunt prelucrate pentru a obţine trăsături (descriptori) de o calitate mai mare.

Introduse original pentru a descrie texturi, Modelele Binare Locale (*Local Binary Patterns* - LBP) [1], [2], au ajuns să fie folosite în mai multe probleme de ML. În varianta

de bază, descriptorii LBP presupun analiza vecinătaţilor V8 din jurul tuturor pixelilor şi marcarea valorilor din acestea faţă de pixelii centrali. După obţinerea codărilor, se realizează o histogramă a lor, care va reprezenta descriptorul final.

Un descriptor de o deosebită popularitate, Histograma Orientărilor de Gradienţi (*Histogram of Oriented Gradients* - HOG) [3] a fost original proiectată pentru detecţia de pietoni. Deşi există mai multe variabile care determină calitatea finală a descriptorului, descriptorii HOG se calculează mereu pe blocuri din imagine.

Descriptorii de tip DeCAF (*Deep Convolutional Activation Feature*) [4] se disting de cei prezentaţi anteriori prin faptul că valorile sunt obţinute prin propagarea imaginilor printr-o arhitectură convoluţională. Plecându-se de la o reţea antrenată pe o problemă generală, descriptorii sunt activările dintr-un strat intermediar, apropiat de cel de ieşire. Abordarea s-a dovedit competitivă, depăsind alte metode în diverse scenarii experimentale.

2.3 Sisteme de învăţare automată

Odată obţinuţi descriptorii folosind metode precum cele amintite anterior, acestea trebuie să fie prelucrate de un sistem de învăţare. În paragrafele următoare vor fi prezentate diverşi algoritmi, cu menţiunea că reţelele neuronale convoluţionale vor fi tratate separat.

Din punct de vedere conceptual, arborii de decizie reprezintă o soluţie simplă pentru învăţarea automată. În fiecare nod al aborelui care nu este de tip frunză, se realizează o partiţie a spaţiului de trăsături, iar răspunsul final este determinat de nodul terminal. Pentru a remedia o parte din problemele de performanţă ale algoritmului, au fost propuse ansamblurile de arbori de decizie (*Random Forests* - RF) [5], în care fiecare structură din componenţă prezintă un grad de aleatorism pentru asigurarea diversităţii.

Scopul unei maşini cu vector suport (*Support Vector Machines* - SVM) este de a găsi hiperplanul optim de separare a datelor. Deşi în varianta lor originală puteau fi folosite doar pentru probleme în care clasele sunt liniar separabile, introducerea *trucului cu nuclee* (sau kernel trick) a permis extinderea funcţionalităţii SVM-urilor.

De-a lungul unei perioade mari de timp, perceptronul multistrat (*Multilayer Perceptron* - MLP) a fost cea mai întâlnită formă de reţea neuronală. Structura de bază a acestuia presupune existenţa unui strat de intrare, a unuia de ieşire şi a cel puţin un strat ascuns. Valoarea fiecărui neuron dintr-un strat este o combinaţie liniară a tuturor neuronilor din stratul anterior, peste care se aplică o *funcţie de activare* care introduce neliniarităţi.

Spre deosebire de celelalte metode menţionate, Algoritmul k-Means [6] se încadrează în categoria neuspervizată. Scopul acestuia este de a grupa datele, pe baza partiţionării spaţiului trăsăturilor. Pentru a realiza acest lucru, etapa de antrenare constă în găsirea celor k centroizi astfel încât să fie minimizate distanţele dintre eşantioane şi cel mai apropiat centroid.

Capitolul 3

Reţele Neuronale Convoluţionale

Cele mai populare soluţii curente de învăţare automată pentru probleme de vedere computerizată sunt Reţelele Neuronale Convoluţionale (*Convolutional Neural Networks* - CNN). Acestea se bazează pe înlănţuirea mai multor tipuri de operaţii (sub forma unor straturi), pentru a prelucra imaginea de intrare şi a obţine răspunsul dorit. În continuare vor fi prezentate principalele straturi utilizate, precum şi o serie de arhitecturi de referinţă.

3.1 Straturile unei reţele neuronale convoluţionale

Deşi nu există o structură fixă pentru diversele arhitecturi, cel mai des sunt întâlnite aceleaşi tipuri de straturi. O serie de studii a condus la diverse îmbunătăţiri, asigurând creşterea performanţei, fără a creşte neapărat numărul de parametri utilizaţi.

Cel mai tipic strat utilizat este cel de convoluţie. Utilizând convoluţia, sunt rezolvate o parte din neajunsurile straturilor folosite în MLP. Utilizând convoluţii, nu mai există conexiuni între toţi neuronii din straturi consecutive, scăzând efortul computaţional. De asemenea, cu ajutorul unui număr redus de ponderi care aparţin unui nucleu poate fi prelucrat tot stratul anterior.

Odată cu avansarea în reţea, poziţiile precise ale unor trăsături devin mai puţin importante, în anumite cazuri cauzând dificultăţi procesului de antrenare. Din acest motiv, majoritatea arhitecturilor utilizează operaţii de subeşantionare, sub forma straturilor de unificare (*Pooling*). Acestea parcurg diversele vecinătăţi din stratul anterior şi înlocuiesc valorile din fiecare vecinătate cu una singură, reprezentând o statistică relevantă a acesteia. Cel mai întâlnit astfel de strat este *max-pooling* care presupune păstrarea valorii maxime.

Pentru a preveni comportamentul nedorit de *memorizare* a setului de date de antrenare, au fost definite straturi speciale de regularizare. Un astfel de strat este cel de tip *Dropout* [7] care inactivează, cu o anumită probabilitate, neuronii din stratul anterior, în timpul antrenării. Alt strat, cel de tip *Batch Normalization* (BN) [8], învaţă statistici despre activările aplicate la intrare şi asigură normalizarea acestora.

3.2 Învăţarea reţelelor neuronale

Procesul de învăţare al CNN-urilor este asemănător cu cel al MLP-urilor. Aplicându-se un pachet de eşantioane la intrarea în reţea, se calculează un cost pe baza predicţiilor şi, pe baza acestora, un *optimizator* calculează modificările aplicate ponderilor. Variaţia parametrilor se face pe baza propagării înapoi a erorii în reţea (*backpropagation*).

Costul obţinut pe baza predicţiilor măsoară calitatea acestora, raportate la etichete, în scenariile supervizate. Pentru sarcini de regresie (în care eticheta este o valoare continuă), costurile folosite cel mai des sunt de tip eroare medie absolută sau eroare pătratică medie. În cazul clasificării însă, cel mai des ieşirea este modelată ca o distribuţie de probabilităţi şi se utilizează entropia încrucişată, raportată la distribuţia ideală, dictată de etichetă.

Pe baza valorii funcţiei cost, optimizatorul ajustează toate ponderile reţelei. Un prim algoritm de tip optimizator este algoritmul de optimizare stochastic bazat pe gradientul negativ (*Stochastic Gradient Descent* - SGD). O atenţie deosebită a fost acordată posibilelor îmbunătăţiri, fapt ce a dus la soluţiile precum Adam [9].

3.3 Arhitecturi neuronale convoluţionale

În prezent se poate identifica un număr semnificativ de arhitecturi care au adus elemente notabile în domeniul CNN-urilor. În continuare vor fi prezentate câteva dintre reţelele cele mai reprezentative din literatură.

Arhitectura AlexNet [10] a impresionat prin performanţele sale la momentul apariţiei. Printre aspectele notabile ale acesteia trebuie amintite folosirea funcţiei de activare ReLU, precum şi utilizarea a două procesare grafice.

După apariţia AlexNet s-a conturat un interes crescut pentru numărul de straturi neuronale prezente în reţea. Arhitectura VGG [11] explorează acest aspect, introducând reţele cu până la 19 straturi. De această dată sunt folosite în mod exclusiv convoluţii cu nucleu 3×3 şi pasul 1, alături de straturi de tip *max-pooling*.

Dimensiunile tot mai mari ale arhitecturilor au devenit o problemă, din cauza dificultăţilor de antrenare cauzate de problema gradientului care dispare. Pentru a remedia această situaţie, arhitecturile de tip *ResNet* [12] propun introducerea unor conexiuni de tip *scurtătură*, care presupun ca intrarea dintr-un bloc de mai multe convoluţii sa fie adunate la ieşirea sa.

Dezvoltând ideea introdusă de ResNet, arhitecturile de tip clepsidră (*Hourglass* - HG) [13] propun o altă grupare a blocurilor convoluţionale. Fiecare modul de tip clepsidră constă într-o zonă de blocuri convoluţionale şi operaţii de tip *max-pooling*, urmate de zone în care este folosită supraeşantionarea pentru revenirea la dimensiunile iniţiale. Cele două tipuri zone sunt apoi conectate cu o legătură precum cea din ResNet.

Capitolul 4

Elemente de învăţare semi-supervizată

Pe lângă tipurile de învăţare menţionate anterior se pot defini şi alte variante, cu diverse utilităţi practice. Îmbinând premisele de bază ale metodelor supervizate şi nesupervizate, învăţarea semi-supervizată (*Semi-supervised Learning* - SSL) presupune utilizarea atât a datelor etichetate cât şi a celor neetichetate pentru a spori performanţa sistemului de bază. Dacă pentru porţiunea supervizată lucrurile sunt clare, provocarea metodelor semi-supervizate constă în găsirea modalităţilor în care pot fi folosite datele fără etichete. O clasificare tipică defineşte metodele de minimizare a entropiei, în care este vizat în mod direct stratul de ieşire, precum şi metodele pe bază de regularizatori [14].

4.1 Condiţii formale de folosire a învăţării semi-supervizate

Pentru a putea utiliza algoritmi SSL, este necesară îndeplinirea mai multor condiţii referitoare la natura datelor de antrenare [15]. În primul rând, dacă două eşantioane X_1 şi X_2 sunt asemănătoare, este de aşteptat ca ele să aparţină aceleiaşi clase (în cazul clasificării). În al doilea rând, zonele de frontieră între clase trebuie să aibă o densitate mică de eşantioane. În al treilea rând, este presupus faptul că spaţiul de dimensionalitate ridicată a eşantioanelor poate fi modelat ca o serie de structuri de dimensionalitate mai mică. În acest caz eşantioanele aparţinând aceleiaşi structuri ar trebui să prezinte şi aceeaşi etichetă [16].

4.2 Exemple de algoritmi

Există o multitudine de algoritmi SSL propuşi în literatură, fiecare cu avantajele şi inovaţiile sale. În continuare, vor fi prezentaţi algoritmi reprezentativi, relevanţi pentru experimentele care vor fi analizate ulterior.

4.2.1 Pseudo-labels

O metodă simplă de utilizare a datelor fără etichete este algoritmul *Pseudo-labels* [17], utilizat pentru sarcini de clasificare. Acesta propune ca eşantioanele neetichetate să fie folosite într-o funcţie cost suplimentară, tot de tip entropie încrucişată. Etichetele necesare sunt calculate pe baza neuronului cu valoare maximă din stratul de ieşire.

4.2.2 Mean Teacher

În timp ce *Pseudo-labels* propune utilizarea datelor fără etichete în acelaşi mod ca cele etichetate, algoritmul *Mean Teacher* [18] merge în altă direcţie. Fiind definite o reţea *student* şi una *profesor* (calculată ca o medie mobilă exponenţială a reţelei student), este introdus un cost suplimentar între cele două. Eşantioanele fără etichete sunt folosite în acest termen, el fiind eroarea pătratică medie dintre predicţiile celor doua reţele.

4.2.3 MixMatch

Structura *MixMatch* [19] este considerabil mai elaborată faţă de algoritmii prezentaţi anterior. Imaginile fără etichete sunt grupate cu cele etichetate şi sunt generate eşantioane noi folosind *MixUp*. În funcţie de tipul imaginii cu influenţă cea mai mare, noile exemple sunt utilizate fie într-un cost de tip entropie încrucişată, fie într-o eroare pătratică medie.

4.3 Tehnici de augmentare a datelor

O modalitate de a creşte performanţa algoritmilor supervizaţi este de a aplica augmentări asupra imaginilor de intrare. Aceste operaţii au însă un impact mult mai mare în algoritmii SSL, iar în continuare vor fi prezentate câteva metode relevante experimentelor.

Cele mai tipice metode de augmentare provin din zona operaţiilor de prelucrare a imaginilor. Utilizarea translaţiilor, a rotaţiilor şi a răsturnărilor sunt frecvent întâlnite în etapa de preprocesare a imaginilor. Suplimentar, se mai întâlnesc augmentări pe baza adăugării de zgomot Gaussian şi chiar a decupărilor din imagini de rezoluţie mai mare decât intrarea.

Algoritmul *MixUp* [20] s-a evidenţiat în ultimii ani, atât pe baza rezultatelor cât şi a simplităţii metodei. Având la dispoziţie două eşantioane, se generează o imagine nouă ca o combinaţie liniară a celor două, cu un factor α. Aceeaşi combinaţie, cu acelaşi α este aplicată şi distribuţiilor descrise de etichetele celor două eşantioane de intrare.

Capitolul 5

Direcţia de cercetare 1: Analiza tablourilor

Problemele de învăţare automată tipice precum detecţia obiectelor presupun cel mai des prelucrarea informaţiei vizuale reprezentate într-o manieră realistă, asemeni situaţiilor reale în care se regăsesc oamenii în fiecare zi. Cu toate acestea, abilitatea umană de a identifica şi a analiza elementele importante dintr-o scenă este mult mai avansată, având abilitatea de a descifra cu uşurinţă şi reprezentări cu un grad mai mare de abstractizare. Această trăsătură a permis dezvoltarea artei vizuale, în care un artist se foloseşte în mod creativ de modalităţi diverse de reprezentare pentru a atrage atenţia publicului sau pentru a transmite mesaje profunde.

Pentru a putea avea rezultate semnificative în domeniul analizei tablourilor, a fost necesară şi existenţa unor colecţii semnificative de artă care să fie disponibile în format digital. Iniţiative precum WikiArt[1] şi Art UK[2] au făcut accesibile sute de mii de opere de artă pentru publicul larg, uşor accesibile pe Internet. Un aspect de menţionat este faptul că fiecare tablou are mai multe etichete ataşate, acolo unde a fost posibil. În cazul lucrării curente, cea mai importantă etichetă considerată a fost cea de *gen* al tabloului. Însemnătatea diverselor genuri nu este întotdeauna evidentă, în unele cazuri referindu-se la subiectul central al tabloului, însă alteori modul de reprezentare este cel care determină genul. Faptul că există o suprapunere între gen şi tipul scenei permite însă considerarea fotografiilor "de consum" ca fiind comparabile cu unele tablouri ca gen.

Etapa experimentală care va fi prezentată în acest capitol urmăreste creşterea performanţei unor arhitecturi convoluţionale în cazul clasificării de gen. După stabilirea unor rezultate de referinţă, au fost utilizate metode de adaptare de domeniu, în principal sub forma transferul de stil, astfel încât să fie extins setul de antrenare iniţial. Problema stilurilor cu grad mare de abstractizare va fi tratată cu o importanţă aparte.

[1] http://wikiart.org
[2] http://artuk.org

5.1 Legătura cu literatura de specialitate

Eforturile de cercetare în zona analizei automate a operelor de artă s-a manifestat dinaintea adoptării Reţelelor Neuronale Convoluţionale ca soluţia ML dominantă, precum în [21], iar în multe cazuri, subiectul de interes este stilul tabloului ([22], [23]). Înaintea introducerii WikiArt, seturile de date utilizate erau considerabil mai mici, cu sub 2000 de eşantioane ([24], [25]).

În anumite cazuri, autorii au abordat atât recunoaşterea genurilor cât şi a stilurilor tablourilor din WikiArt, precum în [26], unde a fost utilizată o arhitectură AlexNet preantrenată. Dintr-o perspectivă adiacentă, trebuie menţionate şi realizări precum cele ale Zhou et al. [27], care şi-au propus recunoaşterea de scene din fotografii, un subiect cu o oarecare apropiere de recunoaşterea de gen.

5.2 Analiza operelor de artă în contextul adaptării de domeniu

O metodă care şi-a demonstrat utilitatea în antrenarea reţelelor neuronale este cea de transfer de domeniu. Prin astfel de abordări se poate realiza creşterea setului de date de antrenare, mecanismele principale fiind descrise de Ben-David et al. [28]. În principal, trebuie definite un domeniu sursă, unul ţintă şi o funcţie de adaptare între cele două. În cazul curent, domeniile sursă au fost diverse seturi de fotografii (artistice sau non-artistice), precum şi un subset al tablourilor din WikiArt. Ţinând cont de natura problemei, funcţiile de adaptare utilizate s-au înscris în categoria celor de transfer de stil.

5.3 Algoritmi de transfer de stil

Metodele care ajustează stilul unei imagini conform unei referinţe există de suficient timp încât să fie consacrate mai multe soluţii cu abordări diferite. Îndepărtându-ne de la algoritmi precum cel al lui Reinhard et al. [29] sau [30], se evidenţiază cei în care se operează pe trăsături dintr-o descompunere pe mai multe nivele ca în transferul de stil Laplacian [31] şi cel neuronal [32].

Primul algoritm de transfer de stil utilizat în lucrare propune utilizarea unei descompuneri piramidale a imaginii pentru a realiza transferul de stil la diverse nivele de detaliu. Se foloseşte filtrarea Laplaciană, iar pe baza valorilor obţinute, se realizează o potrivire a distribuţiilor de gradienţi la fiecare nivel al piramidei. Având imaginea sursă de conţinut C şi cea de stil S, pentru fiecare pixel p, aparţinând vecinătăţii v, la toate

nivelele piramidei se calculează:

$$\begin{aligned} C_n(p) &= r(C_{n-1}(p)) \\ r(i) &= g + sign(i-g)t(|i-g|) \\ t(i) &= F^{-1}_{\nabla S(p)} F_{\nabla C(p)}(p) \end{aligned} \tag{5.1}$$

Algoritmul de transfer de stil utilizând CNN-uri al lui Gatys et al. [32], propune utilizarea trăsăturilor calculate pe parcursul propagării înainte a informaţiei prin reţea în mod separat pentru a descrie stilul sau conţinutul imaginilor. Utilizarea trăsăturilor din straturile inferioare ale reţelei conduc la generarea unei imagini foarte asemănătoare cu cea originală de conţinut. În schimb, dacă sunt folosite straturile superioare, asemănările la nivel de pixel nu mai sunt atât de clare, însă sunt surprinse aranjamentele spaţiale grosiere. În mod similar, pentru stil, straturile inferioare sunt strâns legate de trăsături de stil sau textură foarte fine, iar straturile superioare sunt folositoare pentru a transfera idei de ansamblu.

Procesul efectiv de transfer presupune iniţializarea unei imagini cu zgomot alb X, care este apoi ajustată printr-un proces iterativ astfel încât să semene cu imaginile C şi S. Componentele de stil şi conţinut presupun utilizarea unor funcţii cost separate, adaptate specificului celor două tipuri de imagini de intrare, care sunt apoi reunite într-o combinaţie liniară a costurilor parţiale (5.2).

$$L_{total}(C,S,X) = \alpha L_{continut}(P,X) + \beta * L_{stil}(S,X) \tag{5.2}$$

Asemănarea cu imaginea sursă de conţinut C este realizată minimizând eroarea pătratică între activările obţinute în reţea, aplicând la intrare C şi X (5.3). Elementele de stil utilizează matricile Gram ale activărilor (produsul scalar a două hărţi de activare vectorizate), costul fiind reducerea erorii pătratice între matricile obţinute pentru X şi S la nivele diferite din reţea (5.4).

$$L_{continut}(C,X,l) = \frac{1}{2}\sum_{i,j}(C^l_{ij} - X^l_{ij})^2 \tag{5.3}$$

$$E_l = \frac{1}{4N_l^2 M_l^2}\sum_{i,j}(G^l_{ij} - S^l_{ij})^2 \tag{5.4}$$

Modul de funcţionare al algoritmilor de transfer de stil Laplacian şi neuronal prezintă diferenţe evidente, însă se pot observa şi anumite asemănări. Ambele metode presupun operaţii de filtrare pentru a obţine descriptorii intermediari, însă doar metoda neuronală utilizează filtre învăţate, care extrag informaţii din ce în ce mai complexe, odată cu avansarea în reţea. De asemenea, ambele metode operează la multiple nivele de detaliu, fie prin descompunerea piramidală, fie cu ajutorul straturilor de *pooling* dintr-o arhitectură convoluţională.

5.3.1 Accelerarea transferului de stil neuronal

Pentru a avea un set suplimentar suficient de mare de imagini stilizate este nevoie de timpi mari de procesare, deoarece generarea unui singur pseudo-tablou poate dura până la 20 de minute. Această problemă a fost adresată în [33], fiind investigate câteva posibile soluţii. Una dintre ele a constat în schimbarea arhitecturii neuronale dintr-un VGG-16 sau VGG-19 într-un ResNet-50. După o căutare a combinaţiei optime de straturi, a fost constatat faptul că imaginile rezultate nu aveau aspectul dorit.

O a doua metodă explorată pentru reducerea timpului necesar generării unui număr mare de pseudo-tablouri a constat în găsirea unui moment bun pentru întreruperea timpurie a procesului de optimizare. În urma unei serii de verificări, a fost ales pragul de 100 de iteraţii ca limită superioară de etape de ajustări.

5.4 Seturi de date

În continuare vor fi prezentate seturile de date care au fost considerate pentru experimentele referitoare la clasificarea de gen a tablourilor. Acestea au conţinut fie tablouri, fie fotografii (artistice sau de consum), fie pseudo-tablouri utilizând date din celelalte seturi.

5.4.1 WikiArt

Dimensiunile impresionante (>250,000 de tablouri) precum şi diversitatea operelor conţinute de WikiArt au făcut ca acest set să fie relevant pentru problemele de analiză a tablourilor. Un subset de peste 80,000 de tablouri, utilizat şi în studii trecute precum [23] a fost folosit în experimentele care vor fi prezentate în continuare, având rezultatele raportate în [34] şi [35]. Într-o etapă de verificare a datelor, anumite exemple au fost scoase, din cauza etichetelor lipsă, rămânând 79,434 de imagini utilizabile.

Pentru clasificarea de gen există 42 de cateogrii distincte, însă unele au un număr prea mic de exemple (<200), ceea ce a condus la gruparea acestora într-o clasă separată numită *Altele*. Celelalte 25 de tipuri de gen considerate a fi suficient de bine reprezentate sunt: "Artă Abstractă", "Tablouri alegorice", "Tablouri cu animale", "Tablouri cu bătălii", "Peisaje citadine", "Design", "Tablouri figurative", "Tablouri cu flori", "Tablouri cu viaţa cotidiană", "Tablouri istorice", "Ilustraţii", "Interior", "Peisaje", "Tablouri literare", "Marina", "Tablouri mitologice", "Nuduri", "Portrete", "Poster", "Tablouri religioase", "Autoportrete", "Schiţe şi studii", "Natură moartă", "Tablouri simbolice", "Tablouri cu viaţă sălbatică". O dificultate care a fost observată imediat este suprapunerea între anumite clase, cum ar fi "Portrete" şi "Autoportrete", sau "Tablouri alegorice" şi mai multe alte tipuri de genuri.

5.4.2 Seturi cu fotografii

Pe baza unor asemănări între ideile de *gen* al unui tablou şi scenă, experimentele au utilizat setul *Scene UNderstanding* (SUN) [36] drept o primă sursă suplimentară de date. Din cele 899 de clase prezente în SUN, au fost alese 61 de clase ca fiind de interes imediat pentru îmbunătăţirii clasificării de gen. Ţinând cont de importanţa clasei *Portrete*, au fost utilizate şi imagini din setul *Labeled Faces in the Wild* (LFW) [37].

O componentă importantă în evaluarea oricărui tablou este reprezentată de stilul acestuia, motiv pentru care imaginile din seturi de date precum SUN ar putea fi insuficiente pentru îmbunătăţiri semnificative. Deşi ideea de stil este diferită pentru fotografii decât pentru tablouri, atenţia acordată modului de prezentare a conţinutului ar putea fi importantă pentru problema abordată, motiv pentru care a fost analizată colecţia aleasă de Thomas şi Kovashka în [38]. Plecând de la cele peste 180,000 de fotografii disponibile, au fost alese, în urma unui proces de selecţie manual, aproape 20,000 de exemple adecvate.

5.4.3 Imagini stilizate

Eşantioanele din această categorie reprezintă rezultatul combinării de imagini din celelalte seturi prezentate anterior. Au fost considerate două cazuri principale, pe baza cărora au fost generate noile imagini: transferul de stil de la tablouri la fotografii (utilizând separat transferul de stil Laplacian şi cel neuronal) şi transferul de stil de la tablouri cu grad ridicat de abstractizare la tablouri având conţinutul reprezentat realist.

5.5 Etapa experimentală

Rezultatele raportate pentru clasificarea de gen au utilizat un subset al WikiArt, şi au presupus un set de experimente iniţiale considerate drept referinţă, urmate de testarea diverselor metode de creştere a performenţei de clasificare pe baza augmentării setului de date de antrenare. Suplimentar, a fost analizat şi impactul stilului asupra calităţii predicţiilor reţelei. În studii anterioare precum cele din [39] au fost considerate doar 10 clase, având cea mai bună reprezentare în cadrul setului de date. Din acest motiv a fost realizată o comparaţie iniţială şi cu acest caz. Drept set de testare, au fost păstrate 20% din întregul set de date, alese în mod aleator, raport utilizat şi de Karayev et al. [23], deşi alţi autori au considerat alte proporţii pentru seturile de antrenare şi testare.

Rezultatele de referinţă, precum şi cele obţinute în urma procesului de augmentare cu tehnici clasice sunt raportate în Tabelul 5.1. Se observă că, pe lângă CNN-uri au fost verificate şi metode care presupun o etapă de extragere a trăsăturilor urmată de folosirea unui clasificator. Cele două rezultate obţinute folosind un ResNet-34 şi setul iniţial de date sunt diferenţiate de hiperparametrii de antrenare diferiţi utilizaţi în [34] faţă de [35]. Se poate observa că abordările convoluţionale au condus mereu la obţinerea unor

Soluţie	Număr Clase	Număr imagini	Raport testare	Acurateţe [%]
Saleh şi Elgammal [39] Classemes + Boost	10	63,691	33%	57.87
Saleh şi Elgammal [39] Classemes + ITML				60.28
Tan et. al [26] AlexNet			-	69.29
Tan et. al [26] CNN preantrenat				74.14
ResNet-34 [34]			20%	**75.58**
pLBP + SVM [34]	26	79,434	20%	39.58
pHOG + SVM [34]				44.37
DeCAF + SVM [34]				59.05
AlexNet [34]				53.02
ResNet-34 [35]				59.1
ResNet-34 [34]				**61.64**
ResNet-34 augmentat [34]				**63.58**

Tabel 5.1 Rezultatele clasificării de gen a tablourilor. Sunt comparate experimentele proprii din [34] şi [35] cu cele din literatura de specialitate.

rezultate mai bune. Cel mai bun rezultat a fost obţinut prin creşterea semnificativă a dimensiunii setului de antrenare utilizând tehnicile de simetrizare faţă de axa verticală şi rotirea imaginilor. Această augmentare a condus şi la creşterea semnificativă a timpului de antrenare necesar, motiv pentru care transferul de stil a fost utilizat doar în contextul setului în formatul său iniţial.

5.5.1 Impactul stilului asupra performanţelor de clasificare

Este de aşteptat ca în cazul creşterii gradului de abstractizare a reprezentării conţinutului să fie afectată în mod negativ performanţa reţelei neuronale. Plecând de la această idee, a fost realizată o nouă împărţire a setului de date, astfel încât tablourile din stilurile *Cubist* şi *Artă naivă* să alcătuiască întreg setul de testare (4,132 de tablouri în total). Deşi setul de antrenare a crescut vizibil în acest caz, acurateţea de clasificare a scăzut cu peste 10%, la 50.82%, acurateţea Top-5 micşorându-se şi ea cu aproximativ 8%.

O altă dovadă a importanţei stilului a fost dată de o analiză suplimentară a rezultatelor de bază. Pentru stilurile clasice se pot observa rezultate bune, în timp ce stilurile precum *Artă naivă*, *Cubism* şi *Supraprealism* prezintă valori vizibil scăzute. Excepţiile de la această tendinţă sunt stilurile ca *Minimalism* şi *Color Field Painting*, unde conţinutul abordat este unic.

5.5.2 Transferul de domeniu

Obţinerea imaginilor necesare pentru evaluarea transferului de domeniu a constat în mai multe etape de lungă durată de selecţie sau generare de exemple noi. În final, pentru

acest experiment au fost considerate peste 30,000 de pseudo-tablouri (imagini stilizate cu transferul de stil neuronal) şi peste 20,000 de fotografii artistice. Experimentele au fost realizate în mod iterativ, fiecare iteraţie presupunând un număr maxim de imagini care se poate folosi pentru fiecare gen individual. Din cauza faptului că sursele de imagini suplimentare prezintă un număr diferit de eşantioane pentru diversele clase, numărul de exemple transferate utilizabile a fost fixat pentru fiecare gen. Au fost utilizate astfel 2903 imagini pentru *Peisaje citadine*, 4467 *Peisaje* şi 4002 *Portrete*. Rezultatele raportate în Tabelul 5.2 înfăţişează o serie de aspecte interesante. Pentru început, în toate cazurile, majoritatea performanţei este asigurată de tablourile din WikiArt. Această tendinţă ar putea fi explicată de faptul că tablourile trebuie să aibă o componentă deosebită pentru a fi considerate relevante artistic. De asemenea, utilizarea întregului set de tablouri determină în mod automat eliminarea efectelor benefice ale transferului de domeniu. Pentru a observa îmbunătăţiri, numărul de imagini transferate trebuie să fie peste jumătate din numărul de tablouri utilizate. Un rezultat surprinzător a fost însă reprezentat de faptul că transferul de stil neuronal a avut rezultate mai slabe decât cel Laplacian.

Tip de imagini transferate	Acurateţe raportată la numărul maxim de imagini per clasă					Îmbunătăţire medie
	250	500	1000	5000	Toate	
Fără	19.12	27.95	35.66	54.61	61.64	0
Fotografii naturale	25.28	32.04	38.21	55.31	61.67	2.71
Fotografii artistice	26.72	**32.75**	39.27	53.49	61.55	2.96
Transfer Laplacian pe fotografii artistice	26.56	30.73	39.84	56.17	**61.73**	**3.21**
Transfer neuronal pe tablouri	**26.76**	31.98	37.4	55.33	61.47	2.79
Transfer neuronal pe fotografii artistice	26.33	31.27	**39.94**	54.88	61.62	3.01
Îmbunătăţire maximă	7.64	4.8	4.18	1.56	0.09	—
Imagini adăugate [%]	190.04	103.38	58.98	27.33	≈26	—

Tabel 5.2 Rezultatele experimentelor de transfer de domeniu. Linia marcată cu *Fără* conţine rezultatele în situaţia originală, fără alte imagini adăugate. Coloana *Îmbunătăţire medie* reprezintă media diferenţei de performanţa pentru toate iteraţiile unui singur tip de imagini transferate. Rezultate regăsite în articolul [34].

5.6 Concluzii

Se disting două contribuţii semnificative din experimentele derulate în cadrul acestui capitol. În primul rând, numărul de rezultate este extins, indiferent dacă se face referire la analiza de bază sau cea bazată pe transferul de domeniu. În acest mod au fost subliniate o serie de limitări ale algoritmilor de stil atunci când sunt consideraţi în postura de funcţie de adaptare, în ciuda imaginilor rezultante interesante. În al doilea rând, a fost analizat impactul stilurilor artistice asupra performanţelor de clasificare. Deşi stilurile abstracte sunt problematice pentru arhitecturile convoluţionale, trebuie amintit faptul că şi oamenii au dificultăţi în a analiza astfel de opere de artă.

Capitolul 6

Direcţia de cercetare 2: Analiza facială pentru deducerea emoţiilor

Importanţa comunicării interpersonale este incontestabilă în buna funcţionare a societăţii. Se disting trei componente principale ale comunicării, fiecare fiind abordată de algoritmi ML specifici. Prima componentă este cea verbală, care constă în alegerea cuvintelor folosite pentru a transmite un mesaj. A doua componentă este cea non-verbală care, deşi nu are conţinutul informaţional direct a primei categorii, oferă o informaţii importante asupra stării emoţionale a vorbitorului sau intenţiile sale. Prin natura sa, această componentă necesită folosirea algoritmilor de analiză a informaţiei vizuale. Ultima componentă, cea paraverbală, grupează noţiunile referitoare la modalitatea rostirii cuvintelor: evidenţierea unor cuvinte în discurs, evoluţia toanlităţii vocii, viteza vorbirii, etc.

Capitolul curent se axează pe studierea componentei non-verbale şi are ca scop îmbunătăţirea algoritmilor de analiză a expresiilor faciale. Pentru acest lucru, a fost necesară introducerea unor elemente specifice de ML pentru problema dată, dar şi a unor noţiuni de psihologie. În continuare, vor fi explicate o serie de aspecte teoretice, particularităţi ale folosirii reţelelor neuronale pentru probleme de analiză facială, urmate de prezentarea etapei experimentale.

6.1 Legătura cu literatura de specialitate

Importanţa componentei non-verbale în comunicare este vizibil şi în literatura ştiinţifică, volumul de studii referitoare la analiza expresiilor faciale fiind cu adevărat impresionant. Deşi există mai multe modalităţi în care poate fi tratată problema, cea mai populară variantă constă în împărţirea expresiilor faciale în 6 clase discrete, după cum au fost propuse de Paul Ekman [40]: *fericire*, *tristeţe*, *furie*, *frică*, *surprindere* şi *dezgust*. Suplimentar, este considerată şi expresia *neutră* (lipsa oricărei expresii faciale), iar uneori setul este extins cu expresia *dispreţului*. Un număr semnificativ de metode performante din ultimele decenii pot fi studiate în articolele lui Căleanu [41] şi ale

Sariyanidi et al. [42]. În ultimii ani, soluţiile s-au îndreptat către abordări convoluţionale, precum în Kuo et al. [43], în care autorii au urmărit atât rezultate competitive, cât şi o eficientizare a consumului de memorie şi a puterii de procesare. Mai multe metode bazate pe arhitecturi neuronale pot fi urmărite în studiul lui Li şi Deng [44]. Din punct de vedere al alternativelor la modelul cu şase expresii discrete, se evidenţiază articole precum cel al Zhao et al. [45], care se folosesc de Unităţile de Acţiune Facială (*Action Units* - AU)

Privind în ansamblu soluţiile propuse în literatură pentru evaluarea expresiilor faciale, se poate remarca o serie de tendinţe. Asemeni soluţiilor de recunoaştere facială, pe lângă entropia încrucişată, tipică clasificării, sunt adeseori adăugate şi funcţii cost suplimentare. De asemenea, o metodă de a spori performanţa reţelelor este de a utiliza mai multe tipuri de etichete (expresii şi AU-uri) în timpul antrenării, valorificând legătura dintre acestea.

6.2 Metode de analiză a expresiilor faciale

Atât în ML, cât şi în psihologie, exprimarea obiectivă a expresiilor faciale este un aspect dificil, motiv pentru care, de-a lungul timpului, au fost propuse diverse sisteme de reprezentare a acestora. Prima variantă care trebuie menţionată este cea a lui Ekman [40], care presupune existenţa unei serii de expresii faciale care sunt înfăţişate la fel în orice cultură (enumerate în subcapitolul anterior, reprezentate în Fig. 6.1).

Un sistem prezentat ca o alternativă la cel al expresiilor fundamentale sunt Unităţile de Acţiune Facială (*Action Units* - AU), care propun o exprimare mai granulară şi obiectivă a problemei. Introdus de Ekman şi Friesen [46], acest sistem asociază o serie de coduri diverselor mişcări făcute de muşchii faciali. Deşi sunt definite 27 de AU-uri de acest tip [47], în antrenarea reţelelor neuronale frecvent sunt utilizate doar cele mai comune 10-12 mişcări.

Pe lângă cele două variante de mai sus, s-au definit şi alte metode, precum modelele dimensionale. Acestea înglobează mai multe sisteme în care expresia curentă poate fi reprezentată ca un punct într-un sistem de coordonate, cel mai des bidimensional sau tridimensional. Cea mai comună variantă este modelul bidimensional al lui Russell [48], în care axele reprezintă conceptele de "plăcere" şi "excitaţie" ("activare").

Cea mai populară variantă de analiză a expresiilor faciale rămâne teoria expresiilor fundamentale, deşi toate trei opţiunile prezentate mai sus vin cu seria lor de avantaje şi dezavantaje. Preferinţa pentru acest model poate fi explicată de faptul că este simplu şi permite o obţinere mai puţin costisitoare a etichetelor. În cadrul experimentelor din acest capitol, se vor utiliza atât această variantă, cât şi cea bazată pe recunoaşterea AU-urilor.

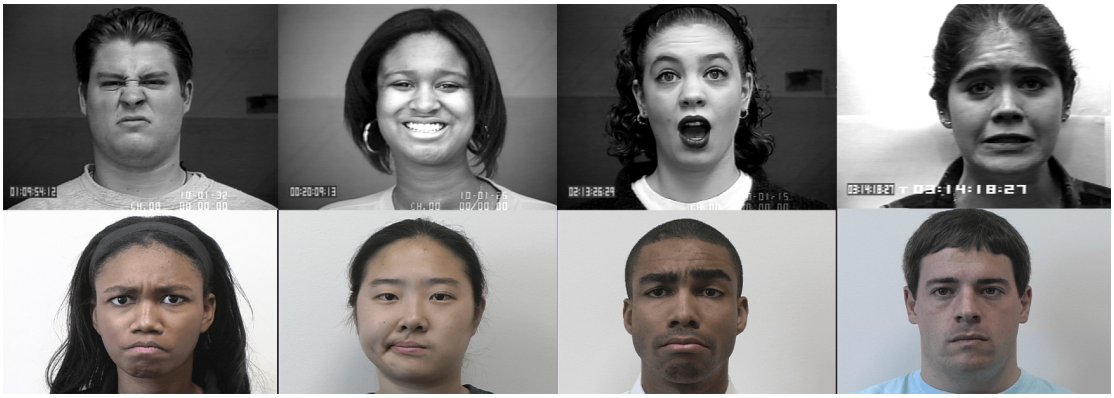

Figura 6.1 Exemple ale celor 7 expresii fundamentale şi cea neutră. De la stânga la dreapta, pe primul rând: *dezgust*, *fericire*, *surprindere*, *frică*. Pe al doilea rând: *furie*, *dispreţ*, *tristeţe*, *neutru*. Imagine preluată din [49].

6.3 Particularităţi ale analizei expresiilor faciale în învăţarea automată

Orice arie de studiu în care se pot folosi algoritmi ML va prezenta o serie de dificultăţi date de particularitatea problemei. Analiza de expresii faciale se poate considera mai dificilă faţă de alte sarcini, precum clasificarea generală a obiectelor, pe baza naturii datelor. În primul rând, alcătuirea imaginilor evidenţiază o diferenţă semnificativă. Dacă pentru seturi precum CIFAR-10 [50] fiecare clasă arată semnificativ diferit faţă de celelalte, seturile cu expresii au acelaşi subiect central: faţa umană. Mai mult, dacă în cazul setului general se poate afirma într-o manieră obiectivă care este eticheta corectă, mereu va exista un grad de subiectivitate în determinarea expresiilor dificile.

În cadrul altei sarcini de analiză facială, recunoaşterea facială, a fost observat faptul că funcţia obişnuită de entropie încrucişată este depăşită de metodele care încurajează generarea de trăsături mai discriminatorii [51]. Această observaţie a condus la introducerea unor funcţii cost noi mai performante, două astfel de funcţii fiind relevante abordărilor propuse în etapa experimentală.

6.4 Metode de preprocesare a imaginilor cu feţe

Etapa de preprocesare are o importanţă deosebită în cazul analizei faciale. Pe lângă operaţii precum normalizarea valorilor, se disting şi procedee specifice domeniului. Probabil cea mai de impact etapă în lanţul de preprocesare este detecţia feţei, care asigură poziţionarea corectă a subiectului în cadru. Suplimentar, se pot face şi ajustări pe baza orientării feţei sau a poziţiei punctelor de interes.

Subiectul detecţiei de feţe este unul dintre cele mai studiate din cadrul algoritmilor de CV. Din acest motiv, au fost propuse numeroase abordări sau îmbunătăţiri soluţiilor existente. Dintre diverşii algoritmi existenţi, sunt amintiţi cel al lui Paul Viola şi Michael

Jones [52], precum şi arhitectura MTCNN (*Multi-Task Cascaded Convolutional Neural Networks*) introdusă de Zhang et al. [53]. Prima metodă a folosit noţiuni precum *imaginea integrală* pentru a asigura viteza de procesare ridicată a unui ansamblu de clasificatori în cascadă. Deşi algoritmul Viola-Jones rămâne o referinţă pentru problemă, soluţiile folosind CNN-uri au ajuns şi în acest caz să fie preferate. În cadrul MTCNN sunt antrenate 3 reţele: P-Net (*Proposal Network*), R-Net (*Refine Network*) şi O-Net (*Output Network*). Ambii algoritmi menţionaţi au fost folosiţi în cadrul experimentelor care vor fi prezentate.

Pe lângă detecţia feţei, un alt pas important în analiza imaginilor faciale este alinierea feţelor (*face alignment*). În acest pas sunt extrase informaţii legate de localizarea punctelor faciale importante pentru a ajusta poziţia şi orientarea feţei. Imaginile cu feţe utilizate în exeprimentele din acest capitol au fost prelucrate ulterior cu ajutorul algoritmului introdus de Kazemi şi Sullivan în [54], implementat în biblioteca DLIB[1].

6.5 Funcţia de cost *Center Loss*

Pentru a încuraja generarea de trăsături discriminative, Wen et al. au introdus un cost nou (*Center Loss*) care poate fi utilizat alături de entropia încrucişată [55]. Funcţia introdusă minimizează distanţele intra-clasă, calculând distanţa între fiecare eşantion x_i şi centroizii claselor de care aparţin c_{y_i} (6.1).

$$L_C = \frac{1}{2}\sum_{i=1}^{N}\|x_i - c_{y_i}\|_2^2 \tag{6.1}$$

6.6 Funcţia de cost *Island Loss*

Costul de tip *Center Loss* şi-a demonstrat experimental utilitatea, însă au fost identificate şi posibile îmbunătăţiri. Introdus de Cai et al. [56], *Island Loss* extinde funcţia prezentată anterior prin adăugarea unui termen care penalizează apropierea oricărei perechi de centroizi ai claselor (6.2).

$$L_{IL} = L_C + \lambda_1 \sum_{c_m \in M} \sum_{\substack{c_n \in M \\ c_m \neq c_n}} \left(\frac{c_m \cdot c_n}{\|c_m\|_2 \|c_n\|_2} + 1 \right) \tag{6.2}$$

6.7 Seturi de date

Varietatea experimentelor care vor fi prezentate a necesitat utilizarea unui număr mare de seturi de date cu diverse tipuri de etichete legate de expresiile faciale. Pe lângă diversele moduri în care pot fi utilizate adnotările, alegerea mai multor seturi a fost

[1] http://dlib.net

necesară deoarece în cazuri precum detecţia de AU-uri, datele sunt considerabil mai greu de obţinut, fiind necesare cursurile de acreditare FACS (*Facial Action Coding System*) pentru adnotarea corectă a imaginilor. Suplimentar, au fost utilizate şi imagini cu feţe care nu prezintă etichete legate de expresiile faciale.

6.7.1 FER+

Setul de date FER+ [57] este alcătuit dintr-un subset al FER2013 [58], adresând anumite probleme întâmpinate în lucrul cu setul original. În varianta iniţială, setul de antrenare a constat în 28709 imagini descărcate de pe Internet, toate exemplele prezentând etichete referitoare la expresia fundamentală prezentă. Pe lângă faptul că o parte din imagini au fost eliminate din colecţia originală, FER+ a presupus şi readnotarea întregului set.

6.7.2 RAF-DB

Prezentat de Li şi Deng în [59], şi setul RAF-DB conţine imagini de pe Internet cu etichete de tip expresii fundamentale. De această dată, cele 29,672 de imagini au fost adnotate direct de un număr mai mare de utilizatori umani, fiecare exemplu fiind evaluat de 40 de persoane.

6.7.3 MegaFace

Scopul MegaFace [60] a fost de pune la dispoziţia comunităţii o colecţie mare de imagini (>1,000,000) cu un număr mare de identităţi unice, pentru a fi folosit în sarcini de recunoaştere facială. Deşi nu prezintă etichete legate de expresiile faciale, numărul crescut de eşantioane a făcut ca acest set să fie de interes pentru experimentele care necesită şi o porţiune neetichetată.

6.7.4 CK+

Spre deosebire de primele două seturi, CK+ (*Extended Cohn-Kanade*) [49] este o referinţă importantă pentru algoritmii care îşi propun detecţia de AU-uri. Cele 593 de secvenţe de imagini înfăţişează 123 de subiecţi cărora li s-a cerut să afişeze diverse expresii faciale, în condiţii de laborator. Fiecare secvenţă pleacă de la expresia neutră, iar ultimul cadru prezintă expresia facială la intensitate maximă.

6.7.5 EmotioNet

Dacă în cazurile anterioare fiecare set prezenta un singur tip de adnotări, EmotioNet [61] conţine informaţii atât despre AU-uri, cât şi despre expresiile discrete prezente în imagini. Numărul mare de exemple (aproximativ 1,000,000) descărcate de pe Internet ar fi reprezentat o dificultate semnificativă pentru etapa de etichetare, motiv pentru care

majoritatea imaginilor au fost adnotate utilizând metode automate performante. Totuşi, pentru a asigura o referinţă puternică, 25,000 de eşantioane au primit etichete de tip AU în mod manual, în vederea competiţiei EmotioNet [62].

6.7.6 UNBC-McMaster Shoulder Pain Expression Archive Database

Un aspect care trebuie luat în considerare atunci când este realizată analiza expresiilor este modalitatea în care ele au fost obţinute. Atunci când sursa de informaţii este Internetul, nu există o garanţie a spontaneităţii expresiilor. Această problemă este rezolvată în setul *UNBC-McMaster Shoulder Pain Expression Archive Database* (UNBC-McMaster), introdus de Lucey et al. în [63]. Cele 48,398 de cadre puse la dispoziţie provin de la 129 de persoane care prezentau dureri în zona umărului. Aceştia au fost rugaţi să facă diverse mişcări cu braţele, cadrele fiind apoi adnotate cu etichete de tip AU-uri.

6.7.7 Expresii faciale la copii

În cadrul experimentelor au fost abordate şi seturi de date în care subiecţii sunt copii. Un prim set, CAFE (*Child Affective Facial Expression*) [64] este format din 1192 de imagini în care 154 de copii cu vârste între 2 şi 8 ani au afişat diverse expresii faciale. Supliemntar, a fost folosit setul LIRIS [65], care conţine 208 secvenţe video în care expresiile a 12 subiecţi cu vârste între 6 şi 12 ani sunt spontane.

6.8 Abordări propuse

În subcapitolele anterioare au fost prezentate o parte din dificultăţile tipice analizei de expresii faciale. Experimentele care vor fi prezentate în continuare au căutat, în principal, să aducă îmbunătăţiri comportamentului reţelelor neuronale convoluţionale utilizând imagini fără etichete.

6.8.1 Funcţia de cost pentru analiza facială folosind pseudo-expresii

Costurile precum *Center Loss* şi *Island Loss* şi-au demonstrat utilitatea în generarea de trăsături discriminative. În [66] am extins conceptul, încurajând comportamentul tipic *Center Loss* într-un scenariu care permite învăţarea semi-supervizată.

Arhitectura de tip AlexNet utilizată a fost antrenată atât pentru recunoaşterea expresiilor fundamentale cât şi a AU-urilor prezente în imagine. Pentru a putea folosi costuri precum *Center Loss* este necesar ca toate clasele să fie mutual exclusive, premisă care nu este îndeplinită atunci când se lucrează cu AU-uri. Din acest motiv au fost definite *pseudo-expresiile*, utilizând corespondeţele între AU-uri şi expresii fundamentale. Astfel,

pentru eşantioanele care au doar acest tip de etichete a fost determinată şi o clasă de tip expresie pe baza combinaţiei de mişcări faciale identificate. Costul total este definit ca:

$$L = \alpha_1 L_S + \alpha_2 L_M \tag{6.3}$$

Pe lângă costul supervizat L_S, este utilizat şi costul suplimentar L_M definit conform (6.4) în [66], unde x_i reprezintă trăsăturile din zona dens conectată a arhitecturii. Pentru cazul utilizării AU-urilor au fost utilizate imagini etichetate din EmotioNet, iar pentru expresii discrete a fost ales RAF-DB. Toate imaginile fără etichete, aproximativ 311,000 au provenit din MegaFace. Rezultatele pe EmotioNet pot fi urmărite în Tabelul 6.1.

$$L_M = \sum_{i=1}^{N} \left(\left\| \frac{x_i}{\|x_i\|_2} - \frac{c^j}{\|c^j\|_2} \right\|_2 - \frac{1}{C-1} \sum_{\substack{k=1 \\ k \neq j}}^{C} \left\| \frac{x_i}{\|x_i\|_2} - \frac{c^k}{\|c^k\|_2} \right\|_2 \right) \tag{6.4}$$

Metodă	Tip	AU_1	AU_2	AU_4	AU_5	AU_6	AU_9	AU_{12}	AU_{17}	AU_{20}	AU_{25}	AU_{26}	AU_{43}	Medie subset	Medie
AlexNet [67]	S	24.2	-	34.7	**39.5**	73.1	-	86.8	-	-	88.5	45.6	-	56.1	-
AlexNet *Center Loss*[55]		**34.4**	30.3	55.3	33.3	69.1	**46.1**	79.3	27.8	**32.3**	84.4	43.2	**48.8**	57.9	48.8
AlexNet WSC[67]	SSL	25.3	-	34.5	39.3	**75.6**	-	**87.4**	-	-	**88.8**	47.4	-	57.0	-
AlexNet *Island Loss*[56]	T	30.4	29.5	**56.7**	30.6	66.7	44.1	77.3	26.7	23.8	83.9	47.3	43.9	56.14	46.7
AlexNet *Large Margin*[66]		34.1	**31.1**	56.6	33.9	71.0	45.1	78.1	**30.9**	25.3	83.8	**50.9**	47.2	**58.33**	**49.0**

Tabel 6.1 Scorul F1 pentru recunoaşterea diverselor AU-uri adnotate manual în EmotioNet. Se diferenţiază cele trei paradigme de antrenare: *Supervizată (S)*, *Semi-supervizată (SSL)* şi *Transfer (T)*. *Medie subset* se referă la medie peste subsetul de AU-uri cele mai comune: 1, 4, 5, 6, 12, 25 şi 26. Rezultate raportate şi în [66].

6.8.2 Îmbunătăţirea reţelelor de dimensiuni mici utilizând SSL

O altă etapă a constat în derularea unei serii de teste, plecând de la premisa îmbunătăţirii performanţelor de clasificare ale reţelelor de dimensiuni mici [68]. În acest scop au fost verificate trei metode, utilizând arhitecturi cu număr redus de straturi convoluţionale (2 sau 3). Prima metodă a presupus o reimplementare modificată a Π-*model* [69]. A doua a introdus un cost suplimentar de reconstrucţie pe baza unei ramuri noi de tip autoencoder în zona dens conectată a reţelei, în timp ce ultima a propus minimizarea distanţei dintre trăsăturile unui eşantion fără etichete şi cel mai apropiat vecin etichetat al său. Peformanţa celor trei metode a fost raportată pe seturile CIFAR-10 şi FER+ (Tabelul 6.2).

Nr. etichete	S +CE	Π-*model*		*Cost autoencoder*		*Minimizarea distanţei NN*
		SSL	SSL+ 3 aug.	S	SSL	SSL
320	50.44	51.88	51.31	52.38	52.44	52.92
400	49.16	50.38	49.7	51.49	51.34	53.58
2000	57.91	60.77	60.17	63.67	63.46	64.29
4000	66.2	65.04	64.85	67.25	68.53	70.47
10000	71.82	72.92	72.89	74.11	74.26	76.59

Tabel 6.2 Acurateţea de clasificare pe setul FER+ în cazul experimentelor care utilizează arhitecturi de dimensiuni mici şi tehnici SSL, exprimată în procente. Coloanele care conţin *S* reprezinta antrenări unde nu au fost folosite imagini suplimentare. Coloanele care conţin *SSL* au folosit şi restul de imagini din setul de date, dar fară etichete. Cazul fară costuri suplimentare a fost marcat cu *CE*. Rezultate raportate şi în [68]

.

6.8.3 Margin-mix

În timp ce primele două metode utilizează direct imagini fără etichete, algoritmul Margin-mix, introdus în [70] propune o altă abordare, în care aceste eşantioane sunt combinate cu imaginile din setul etichetat, utilizând MixUp [20]. Metoda necesită etichete pentru generarea de date, motiv pentru care a fost folosită o soluţie asemănătoare *Pseudo-Labels*, dar în spaţiul trăsăturilor, combinată cu utilizarea costului de tip *Large Margin Loss*. În final, costurile suplimentare, utilizate pe lângă entropia încrucişată au fost de tip *Large Margin Loss*, diferenţiate în funcţie de tipul de eşantion utilizat. Pentru partea de analiză a expresiilor, rezultatele pe setul FER+ confirmă calitatea metodei (Tabelul 6.3).

Algoritm \ Nr. etichete	320	400	2000	4000	10000	Toate
WideResNet-28-2	nc	37.92	50.29	56.78	63.56	84.88
Supervizat [57]	-	-	-	-	-	84.99
MeanTeacher [18]	-	45.56	50.84	58.28	68.36	-
MixMatch [19]	45.60	50.25	58.35	70.91	71.24	-
Margin-mix [70]	50.76	56.75	60.83	75.18	81.25	85.36

Tabel 6.3 Comparaţie între diverşi algoritmi folosiţi împreună cu arhitectura WideResNet-28-2, pe setul FER+. Cazurile în care nu a existat convergenţă a algoritmului de antrenare sunt marcate cu "nc". Rezultate raportate şi în [70].

Generalitatea metodei a fost confirmată testând-o mai întâi pe seturi generale precum CIFAR-10, CIFAR-100 [50], SVHN [71] şi STL-10 [72].

6.8.4 Alte experimente

Pe lângă abordările prezentate mai sus, metoda propusă în [73] tratează o problemă importantă în utilizarea metodelor de tip *Pseudo-Labels* pentru probleme de învăţare

semi-supervizată. Algoritmul de etichetare, deşi simplu, poate fi ineficient dacă distribuţiile seturilor de date utilizate sunt puternic diferenţiate. În [73] s-a utilizat o tehnică suplimentară de regularizare pentru a contracara această problemă.

Datele fără adnotări sunt etichetate în mod automat cu *Pseudo-Labels*, însă ponderile reţelei sunt ajustate doar dacă îmbunătăţirile trec de un prag minim determinat de o funcţie aplicată unui parametru de tip temperatură. Suplimentar, se introduc perturbaţii aleatoare costului obţinut pentru eşantioanele neetichetate. Acestea au fost considerate tot din setul MegaFace. În Tabelul 6.4 sunt raportate rezultatele pe setul FER+, alte teste fiind derulate pe RAF-DB. Suplimentar, au fost considerate alte două situaţii. Prima dată, analiza a fost extinsă la seturi cu expresii faciale în care subiecţii sunt copii, în timp ce a a doua oară a fost introdusă o expresie nouă, a *anxietăţii* în setul RAF-DB, împreună cu o serie de imagini relevante.

Metodă	Tip	FER2013	FER+
AlexNet [74]	S	61.10	-
AlexNet [73]		68.2	78.08
FSN [74]		67.60	-
VGG-13 (vot majoritar) [57]		-	83.85
VGG-13 (eşantionarea etichetei) [57]		-	84.99
FUS [75]		-	67.03
AlexNet [73] + *Pseudo-Labels*	T	69.12	80.05
AlexNet + *Pseudo-Labels* + ALRAO [73]		68.65	80.60
AlexNet + ALT [73]		69.62	82.38
VGG-16 + *Pseudo-Labels* [73]		69.27	84.35
VGG-16 + *Pseudo-Labels* + ALRAO [73]		69.27	82.15
VGG-16 + ALT [73]		69.85	85.20

Tabel 6.4 Acurateţea de clasificare pentru problema expresiilor fundamentale pe setul FER şi pe setul FER+. Cu *ALT* este notată metoda de transfer de etichete folosind călire simulată (*Annealed Label Transfer*) din [73]. Cele două metode din [57] se diferenţiază prin moduri diferite de obţinere a etichetelor. Rezultate raportate şi în [73].

6.9 Concluzii

În cadrul acestui capitol au fost propuse diverse metode de sporire a performanţei CNN-urilor utilizate pentru analiza expresiilor faciale. Ţinând cont de diversele moduri de abordare tipice, a fost considerat atât cazul expresiilor fundamentale cât şi cel al AU-urilor. Un aspect important care trebuie subliniat este că metodele s-au remarcat prin integrarea ideilor specifice învăţării semi-supervizate.

Capitolul 7

Concluzii

Deşi au trecut aproape 10 ani de când CNN-urile au început să atragă atenţie, ritmul de apariţie al inovaţiilor a rămas ridicat în ultima perioadă. Progresul este vizibil atât în contextul general al arhitecturilor şi altor elemente din procesul de antrenare, cât şi pentru aplicaţii specifice unui anume domeniu de probleme practice. În lucrarea de faţă au fost alese două tematici diferite pe baza cărora să fie discutate metodele de îmbunătăţire a performanţelor: analiza tablourilor şi cea a expresiilor faciale.

Învăţarea automată s-a dezvoltat în multiple direcţii de când au fost puse bazele domeniului, fiecare cu viziunea ei unică asupra problemei. Multe dintre metodele anterior de vârf prezentate în Capitolul 2, au pierdut teren în faţa abordărilor neuronale actuale. Ele pot avea însă rezultate competitive şi reprezintă un etalon bun în multe situaţii. Deşi anumite aspecte generale despre reţelele neuronale se regăsesc aici, din cauza importanţei CNN-urilor în experimentele prezentate, Capitolul 3 a fost dedicat acestora, completând şi anumite aspecte neacoperite din porţiunea anterioară.

Pe lângă tematicile precise în care au fost folosite reţelele, un interes special a fost acordat învăţării semi-supervizate. În Capitolul 4 au fost prezentate bazele necesare pentru aplicaţiile propuse pe parcursul lucrării. S-a putut observa abilitatea unor algoritmi de a spori performanţele folosind imagini fără etichete. Potenţialul viitor este vizibil, iar algoritmii SSL ar putea adresa scenarii mult mai generale, odată cu dezvoltarea lor.

7.1 Rezultate obţinute

Rezultatele prezentate pe parcursul lucrării reprezintă rezultatul mai multor ani de cercetare în domeniul CNN-urilor. Iniţial axate pe soluţii din zona medicală ([76], [77]), eforturile de cercetare au fost apoi urmate de aplicaţiile cu un volum de date disponibil mai mare, detaliate în continuare.

Capitolul 5 a introdus discuţia legată de clasificarea genului unui tablou. Natura cu totul diferită faţă de restul lucrării până în acel punct a cerut o introducere separată, care să prezinte domeniul, înainte de a aborda elementele de învăţare automată. Pe

lângă reţelele neuronale convoluţionale, într-un rol central în această porţiune a fost ideea de transfer de stil. Au fost prezentate şi utilizate două soluţii pentru generarea de noi imagini care să fie folosite alături de tablouri. Din punct de vedere al datelor folosite, pe lângă seturile de date cu tablouri, au fost folosite şi unele formate din imagini fotografice, dintre care unul cu fotografii artistice. S-a observat că utilizarea transferului de stil poate creşte acurateţea de clasificare, însă problema tablourilor cu un grad mare de abstractizare rămâne o provocare greu de adresat.

Cealaltă mare zonă de aplicaţii care a fost tratată în detaliu este cea a analizei expresiilor faciale, utilizând reţele neuronale convoluţionale. În capitolul 6, s-au discutat aspecte legate de psihologie care să explice formalismele utilizate în continuare. Un accent deosebit a fost pus pe expresiile faciale fundamentale şi Unităţile de Acţiune Facială, care au fost folosite, şi separat, şi împreună, în etapa experimentală. De asemenea, din cauza conţinutului specific al imaginilor faciale, a fost necesară şi prezentarea unor funcţii cost speciale. Pe baza acestora au fost construite şi funcţiile de cost noi din experimentele prezentate, utilizând şi elemente de învăţare semi-supervizată. S-a putut observa îmbunătăţirea rezultatelor, confirmând eficacitatea soluţiilor propuse, precum şi potenţialul abordărilor SSL din această sferă.

7.2 Contribuţii originale

- A fost realizată o analiză extinsă a literaturii de specialitate atât pentru abordările neuronale cât şi pentru cele non-neuronale, pentru domeniile clasificării de gen al tablourilor şi analizei de expresii faciale. Pentru completarea comparaţiilor au fost derulate experimente suplimentare, în funcţie de necesitate ([35], [34]).

- Din surse disponibile, au fost create două seturi auxiliare de imagini adecvate pentru augmentarea colecţiei de tablouri în vederea clasificării de gen. Cele două seturi se disting prin scopul iniţial al surselor, una dintre acestea fiind construită în mod specific pentru a conţine fotografii cu valoare artistică ([34]).

- După o comparaţie conceptuală a două metode de transfer de stil, amândouă au fost utilizate pentru a genera eşantioane noi care să fie introduse în procesul de antrenare. Pe lângă creşterea efectivă a dimensiunii setului de date, a fost vizată şi o creştere a diversităţii modului de reprezentare, prin transferul de stiluri moderne ([34]).

- Pentru adresarea timpilor crescuţi de generare de eşantioane noi, au fost analizate soluţii pentru accelerarea procesului ([33]).

- O analiză a impactului stilului unui tablou asupra acurateţii de clasificare a genului a fost realizată, făcând mai clară dificultatea introducerii stilurilor cu un grad crescut de abstractizare ([34]).

- Au fost propuse multiple metode de îmbunătăţire a acurateţii de clasificare a expresiilor faciale. O componentă specială a acestei porţiuni este reprezentată de integrarea noţiunilor de învăţare semi-supervizată. O primă serie de experimente a vizat utilizarea legăturii dintre codificarea mişcărilor muşchilor faciali şi expresiile fundamentale. Rezultatul a fost o reţea neuronală care a fost antrenată să prezică etichete pentru ambele variante de analiză a expresiilor, utilizând şi eşantioane nesupervizate în contextul unui cost suplimentar pentru sporirea performanţei ([66]).

- O a doua soluţie pentru recunoaşterea expresiilor faciale a fost propusă, aceasta axându-se pe utilizarea unei metode noi de regularizare bazate pe ideea de călire simulată. Suplimentar, au fost investigate şi cazul particular al expresiilor faciale la copii, precum şi cel al expresiei de *anxietate* ([73]).

- În al treilea set de experimente a fost introdus un algoritm care îmbină utilizarea atât de eşantioane cu etichete, cât şi fără etichete, specifică învăţării semi-supervizate, cu o tehnică suplimentară de sporire a numărului de imagini folosite la antrenare. Acest algoritm a fost testat atât în contextul expresiilor faciale, cât şi în cel al unor seturi de date populare de uz general ([70]).

- Pe lângă cei trei algoritmi menţionaţi anterior, au fost desfăşurate o serie de alte experimente la scară mai mică, utilizând arhitecturi de dimensiuni reduse, care să integreze concepte de învăţare semi-supervizată pentru sporirea performanţelor de clasificare a expresiilor fundamentale. Pentru a testa potenţialul de a fi utilizate şi pentru alte sarcini, acestea au fost testate şi pe un set de date clasic de clasificare.

7.3 Lista lucrărilor originale

- Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Efficient domain adaptation for painting theme recognition. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2017

- Corneliu Florea, Mihai Badea, Laura Florea, and Constantin Vertan. Domain transfer for delving into deep networks capacity to de-abstract art. In *Scandinavian Conference on Image Analysis*, pages 337–349. Springer, 2017

- Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Can we teach computers to understand art? domain adaptation for enhancing deep networks capacity to de-abstract art. *Image and Vision Computing*, 77:21–32, 2018

- Andrei Racoviteanu, Mihai-Sorin Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Large margin loss for learning facial movements from pseudo-emotions. In *BMVC*, page 108, 2019

- Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 1–17. Springer, 2020

- Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *BMVC*, page 104, 2019

- Andrei Racoviţeanu, Corneliu Florea, Mihai Badea, and Constantin Vertan. Spontaneous emotion detection by combined learned and fixed descriptors. In *2019 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2019

- Andrei Racovițeanu, Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Dual task training for face expression recognition. In *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE, 2020

- Andrei Racoviţeanu, Iulian Felea, Laura Florea, Mihai Badea, and Corneliu Florea. Clustering based reference normal pose for improved expression recognition. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 51–61. Springer, 2018

- Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Automatic burn area identification in color images. In *2016 International Conference on Communications (COMM)*, pages 65–68. IEEE, 2016

- Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Severe burns assessment by joint color-thermal imagery and ensemble methods. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–5. IEEE, 2016

- Mihai Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Andrei Racoviteanu. Improving small convolutional neural networks with semi-supervised learning. Submitted to UPB Scientific Bulletin, Series C: Electrical Engineering

- Proiectul *Evaluarea Arsurilor prin Imagistică Multi-Spectrală - BAMSI*, PN-II-PT-PCCA-2013-4-0357

- Proiectul *Analiza şi descrierea perceptuală a artei vizuale româneşti - PANDORA*, PN-II-RU-TE-2014-4-0733

- Proiectul *Tehnologii şi sisteme video/audio inovative pentru recunoaşterea/identificarea persoanelor şi a comportamentului simulat - SPIA-VA*, PN-III-P2-2.1-SOL-2016-02-0002

7.4 Perspective de dezvoltare ulterioară

Indiferent de perspectiva avută în vedere, anume cea a aplicaţiilor specializate (analiza tablourilor sau analiza expresiilor faciale) sau cea a tehnicilor de augmentare a procesului de învăţare, se poate afirma că posibilităţile de dezvoltare sunt numeroase. Probabil cea mai importantă direcţie de urmărit este cea a învăţării semi-supervizate. Deşi au fost abordate mai multe soluţii originale, precum şi unele din literatură, s-au observat o serie de probleme generale.

Majoritatea tehnicilor semi-supervizate arată un impact semnificativ pentru seturi de date mici, însă în contextul problemelor uzuale cu seturi mari de date, importanţa lor nu este garantată. O posibilă explicaţie pentru acest comportament este legat de raportul dintre numărul de imagini cu etichete şi cele fără etichete. Dacă luăm în considerare acest aspect, algoritmii semi-supervizaţi ar putea beneficia semnificativ din augmentarea colecţiei de date prin tehnici generative şi soluţii precum transferul de stil, chiar şi în afara contextului artistic al acestuia.

Plecând de la ideea intuitivă menţionată anterior se poate extinde interacţiunea între componenta de cost semi-supervizat şi cea de proces generativ. Procesul de generare ar putea fi controlat în timpul antrenării pe baza costului componentei semi-supervizate, în funcţie de starea curentă a reţelei. Ţinând cont de potenţialul vizibil al învăţării semi-supervizate, putem considera domeniul CNN-urilor ca având înca loc de creştere în zone de aplicaţii în care disponibilitatea datelor este încă un impediment.

Bibliografie

[1] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[2] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR5)*, volume 1, pages 886–893. IEEE, 2005.

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[14] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3239–3250, 2018.

[15] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning*. The MIT Press, 2006.

[16] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[17] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, 2013.

[18] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1195–1204, 2017.

[19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[21] A Bentkowska-Kafel and J Coddington. Computer vision and image analysis of art. In *Proceedings of the SPIE Electronic Imaging Symposium*, 2010.

[22] Corneliu Florea, Cosmin Toca, and Fabian Gieseke. Artistic movement recognition by boosted fusion of color structure and topographic description. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 569–577. IEEE, 2017.

[23] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference. BMVA Press*, 2014.

[24] Siddharth Agarwal, Harish Karnick, Nirmal Pant, and Urvesh Patel. Genre and style based painting classification. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 588–594. IEEE, 2015.

[25] Razvan George Condorovici, Corneliu Florea, and Constantin Vertan. Painting scene recognition using homogenous shapes. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 262–273. Springer, 2013.

[26] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Ceci nst pas une pipe: A deep convolutional network for fine-art paintings classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3703–3707. IEEE, 2016.

[27] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[28] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[29] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[30] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001.

[31] Mathieu Aubry, Sylvain Paris, Samuel W Hasinoff, Jan Kautz, and Frédo Durand. Fast local laplacian filters: Theory and applications. *ACM Transactions on Graphics (TOG)*, 33(5):1–14, 2014.

[32] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[33] Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Efficient domain adaptation for painting theme recognition. In *2017 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2017.

[34] Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Can we teach computers to understand art? domain adaptation for enhancing deep networks capacity to de-abstract art. *Image and Vision Computing*, 77:21–32, 2018.

[35] Corneliu Florea, Mihai Badea, Laura Florea, and Constantin Vertan. Domain transfer for delving into deep networks capacity to de-abstract art. In *Scandinavian Conference on Image Analysis*, pages 337–349. Springer, 2017.

[36] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[37] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[38] Christopher Thomas and Adriana Kovashka. Seeing behind the camera: Identifying the authorship of a photograph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3494–3502, 2016.

[39] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, (2), 2016.

[40] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98:45–60, 1999.

[41] Cătălin-Daniel Căleanu. Face expression recognition: A brief overview of the last decade. In *2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 157–161. IEEE, 2013.

[42] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2014.

[43] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2121–2129, 2018.

[44] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[45] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[46] Paul Ekman and Wallace V Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[47] Emily B Prince, Katherine B Martin, Daniel S Messinger, and M Allen. Facial action coding system, 2015.

[48] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[49] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 94–101. IEEE, 2010.

[50] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[51] Yash Srivastava, Vaishnav Murali, and Shiv Ram Dubey. A performance evaluation of loss functions for deep face recognition. In *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pages 322–332. Springer, 2019.

[52] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[54] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[55] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[56] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James Oeilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[57] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016.

[58] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[59] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018.

[60] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. 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, 2016.

[61] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[62] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*, 2017.

[63] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 57–64. IEEE, 2011.

[64] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, 5:1532, 2015.

[65] Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer, and Saida Bouakaz. A novel database of childrenş spontaneous facial expressions (liris-cse). *Image and Vision Computing*, 83:61–69, 2019.

[66] Andrei Racoviteanu, Mihai-Sorin Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Large margin loss for learning facial movements from pseudo-emotions. In *BMVC*, page 108, 2019.

[67] Kaili Zhao, Wen-Sheng Chu, and Aleix M Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2090–2099, 2018.

[68] Mihai Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Andrei Racoviteanu. Improving small convolutional neural networks with semi-supervised learning. Submitted to UPB Scientific Bulletin, Series C: Electrical Engineering.

[69] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. 2017.

[70] Corneliu Florea, Mihai Badea, Laura Florea, Andrei Racoviteanu, and Constantin Vertan. Margin-mix: Semi-supervised learning for face expression recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 1–17. Springer, 2020.

[71] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[72] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[73] Corneliu Florea, Laura Florea, Mihai-Sorin Badea, Constantin Vertan, and Andrei Racoviteanu. Annealed label transfer for face expression recognition. In *BMVC*, page 104, 2019.

[74] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, page 317, 2018.

[75] Elizabeth Tran, Michael B Mayhew, Hyojin Kim, Piyush Karande, and Alan D Kaplan. Facial expression recognition using a large out-of-context dataset. In *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 52–59. IEEE, 2018.

[76] Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Automatic burn area identification in color images. In *2016 International Conference on Communications (COMM)*, pages 65–68. IEEE, 2016.

[77] Mihai-Sorin Badea, Constantin Vertan, Corneliu Florea, Laura Florea, and Silviu Bădoiu. Severe burns assessment by joint color-thermal imagery and ensemble methods. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–5. IEEE, 2016.

[78] Andrei Racovițeanu, Corneliu Florea, Mihai Badea, and Constantin Vertan. Spontaneous emotion detection by combined learned and fixed descriptors. In *2019 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE, 2019.

[79] Andrei Racovițeanu, Mihai Badea, Corneliu Florea, Laura Florea, and Constantin Vertan. Dual task training for face expression recognition. In *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4. IEEE, 2020.

[80] Andrei Racoviţeanu, Iulian Felea, Laura Florea, Mihai Badea, and Corneliu Florea. Clustering based reference normal pose for improved expression recognition. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 51–61. Springer, 2018.