# Detecția emoțiilor faciale ale preșcolarilor

Radu Corcan, Iulia Strîmbu, Beniamin Pantea

## 1 Obiectiv

Dezvoltarea unei aplicații de învățare pentru copii de vârstă preșcolară. Componenta inteligentă are ca scop recunoașterea emoțiilor copiilor în timp ce aceștia utilizează produsul, obținând informații (feedback) pe baza imaginilor, ce pot fi apoi interpretate și folosite pentru îmbunătățirea aplicației. De asemenea, va exista și un sistem inteligent de autentificare a preșcolarilor, bazat pe recunoaștere facială.

Din punct de vedere formal, urmărim dezvoltarea unui algoritm de învățare supervizată ce rezolvă o problemă de clasificare multi-clasă (împărțirea emoțiilor în categorii: fericit, trist, curios, etc), cu ajutorul rețelelor neuronale convoluționale (deep learning).

## 2 Funcționalități

Preșcolarul se autentifică pe bază de recunoaștere facială, după care este înregistrat (video) în timp ce utilizează soft-ul educațional. Apoi, persoanele interesate (dezvoltatori, educatoare, psihologi, părinți, etc) vor putea vedea clasificarea emoțiilor predominante ale copilului, pe baza predicției algoritmului inteligent. Aplicația va afișa atât emoția în timp real, dar se vor putea genera și două rapoarte: un pie chart ce surprinde proporția fiecărei emoții și un heat map ce redă corelația dintre perechi de emoții (cu alte cuvinte, cât e de probabil ca algoritmul să "încurce" cele două emoții).

## 3 Abordări înrudite

***1. Eduard Franți et al., Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots, Romanian Journal of Information Science and Technology, 2017, 222-240***

Lucrarea tratează recunoașterea emoțiilor pe baza vocii, propunându-și să folosească această clasificare la îmbunătățirea roboților-asistenți personali.

În prima parte, sunt descrise atât importanța detectării emoțiilor în interacțiunea om-calculator, cât și direcțiile de cercetare de până în acest moment. În sens larg, identificarea emoțiilor utilizatorilor unui produs software poate fi folosită pentru îmbunătățirea acestuia, iar cu cât o aplicație "înțelege" mai bine trăirile

persoanei care interacționează cu acesta, cu atât va putea să ofere o experiență mai plăcută. De altfel, s-a observat că oamenii tind să răspundă emoțional la feedback-ul unui calculator (de exemplu, mesaje de eroare) ca și cum acesta ar fi o altă persoană. Autorii argumentează că detecția bazată pe voce este mai precisă, deoarece mimica este mai ușor de controlat (și emoțiile faciale mai ușor de falsificat sau ascuns) decât tonul și inflexiunile vocii.

În ceea ce privește clasificarea emoțiilor umane, autorii se folosesc de taxonomia psihologului american Robert Plutchik, acesta dezvoltând o așa-numită "roată a emoțiilor", ce combină opt sentimente fundamentale (diferențiate fiecare în funcție de intensitate, pe trei niveluri) în perechi de emoții primare și secundare, rezultând în total 32 de emoții. S-a observat însă că nu toate acestea sunt la fel de relevante în interacțiunea om-calculator, astfel că s-a ajuns la un model simplificat, utilizat la scară largă în recunoașterea automată a emoțiilor, ce conține șase categorii: fericit, trist, dezgustat, furios, speriat, surprins. Ar mai fi de menționat totuși că rareori se poate identifica o singură emoție în stare pură; cel mai adesea, se observă combinații, în diferite grade, de cel puțin două trăiri.

Referitor la arhitectura algoritmului inteligent folosit de autori pentru rezolvarea problemei, aceștia se îndreaptă spre rețele neuronale convoluționale (deep learning). Rețeaua folosită are 20 de straturi de convoluție (dimensiune kernel 5x5, funcție de activare ReLU) și unul de max-pooling (dimensiune 2x2), urmat de un strat de flattening, ce generează input-ul pentru o rețea fully-connected cu 1000 de neuroni pe stratul ascuns (activare ReLU). Pe ultimul strat se află clasificatorul cu șase clase (softmax). Autorii obțin astfel o acuratețe de aproximativ 71%, despre care spun că este comparabilă cu acuratețea în detecția emoțiilor pe baza imaginilor faciale, conform unor rezultate obținute anterior. Baza de date de antrenament este relativ mică însă, de doar 200 de fișiere audio, astfel că autorii propun folosirea unui set de date mai mare pentru obținerea de rezultate mai bune, împreună cu analizarea lexicală a înregistrărilor de voce, în paralel cu procesarea semnalului audio.

***2. Poonam Dhankhar, ResNet-50 and VGG-16 for recognizing Facial Emotions, International Journal of Innovations in Engineering and Technology (IJIET)***

Lucrarea realizează în prima parte o trecere în revistă a arhitecturilor de rețele convoluționale menționate și în titlu, prezentând pe scurt modul lor de funcționare, hiperparametri și rezultate obținute pe seturile de date FER 2013 și KDEF [2]. În continuare, autorii încearcă să "combine" modelele învățate de rețelele VGG-16 și ResNet-50, concatenând vectorii de ponderi de pe penultimul strat din fiecare rețea, iar vectorul obținut este considerat input-ul pentru un model de regresie logistică (clasificare). Se obține astfel o îmbunătățire de aproximativ două procente a acurateții față de folosirea celor două rețele în mod individual, iar utilizarea *transfer learning* (se antrenează modelul pe setul de date FER, se fixează ponderile rețelei în afară de ultimele câteva straturi și se continuă antrenarea pe celălalt set de date) conduce și ea la o creștere a acurateții de 2.5%.

***3. Byoung Chul Ko, A Brief Review of Facial Emotion Recogni-***

***tion Based on Visual Information***

Acest articol, de dimensiuni simțitor mai mari decât celelalte două amintite anterior, își propune nu atât o prezentare a unei inovații foarte țintite, de natură tehnică (ce să se concentreze pe un aspect foarte specific din domeniul recunoașterii automate de emoții), ci mai degrabă o expunere a acestui domeniu în sens larg. Sunt prezentate, în ordine cronologică, abordările folosite, de la detecție de fețe și emoții folosind diverse metode "manuale" (cum ar fi filtre Gabor, filtre Haar, histograma gradienților, SIFT), trecând prin clasificatori ce nu țin de zona de deep learning (AdaBoost — similar ca abordare cu clasificatorii în cascadă (mai mulți clasificatori "slabi" ce reușesc împreună să clasifice corect un input) — și mașini cu suport vectorial), până la metodele de deep learning cele mai recente (CNN: Inception, ResNet, etc). Sunt discutate apoi caracteristici ale diverselor baze de date pentru detecție de emoții — ce conțin atât imagini individuale, cât și video-uri —, cum ar fi tipul imaginii (2D, 3D, infraroșu), varietatea persoanelor din imagini (rasă, vârstă), calitatea imaginilor, etc. Mai sunt prezentate și diferitele metrici de performanță a modelelor: acuratețe, precizie, rapel (recall), F-score.

Dincolo de această privire de ansamblu (ce oferă un punct bun de pornire în înțelegerea subiectului, a provocărilor pe care le ridică și a demersurilor de până acum), am considerat interesant în particular o prezentare a detecției de emoție în video-uri (sau secvențe de câteva imagini ce se succed temporal) folosind o combinație între rețele convoluționale (folosite pentru a genera features) și o rețea recurentă (de tip LSTM), ce încearcă să utilizeze aspectul temporal pentru a prezice mai bine o emoție (de exemplu, se poate împărți o emoție în început, sfârșit și parte de mijloc, iar modelul va învăța să folosească succesiunea lor).

# 4 Rezultate obținute

## 4.1 Recunoaștere facială

Pentru prima parte a aplicației, am folosit un model [5] antrenat pe baza de date *Labeled Faces in the Wild* (de referință în domeniul recunoașterii faciale; conține peste 13.000 de imagini), pe care s-a obținut, conform autorilor, o acuratețe de 99.38%. Prezentăm în continuare, pe scurt, pașii făcuți pentru recunoașterea facială, așa cum sunt ei descriși de autorii instrumentului [3]:

1. Se transformă imaginea în alb-negru și se creează histograma gradienților orientați (HOG — o imagine formată din vectori orientați în direcția schimbării luminozității), folosită pentru detecția feței în imagine.

2. Se identifică anumite puncte de pe față (landmarks) și se aplică transformări (afine) asupra imaginii pentru a centra fața.

3. Se antrenează o rețea convoluțională ce generează pentru fiecare față un vector de 128 de valori reale (*embedding*), astfel încât pentru două imagini diferite ale aceleiași persoane vor fi generați doi vectori foarte asemănători, iar pentru persoane diferite — vectori diferiți. Antrenarea rețelei se face

primind două imagini ale unei persoane cunoscute, și o a treia imagine a uneia necunoscute, iar obiectivul este ca diferența dintre distanța euclidiană dintre vectorii generați pentru cele două persoane cunoscute și cea dintre o persoană cunoscută și una necunoscută să fie maximă.

4. Se folosește o mașină cu suport vectorial pentru a clasifica o nouă imagine într-una dintre clasele pe care le cunoaște (fiecare persoană e o clasă, plus o clasă pentru necunoscuți).

#### 4.1.1 Descrierea setului de date

Setul de date folosit (în cazul ambelor modele) este cel utilizat în cadrul la *Facial Emotion Recognition (FER) Challenge - 2013* [1]. Acesta este compus din aproximativ 28000 de imagini pentru setul de antrenament și câte circa 3500 de imagini pentru seturile de validare, respectiv testare. Imaginile sunt alb-negru și au dimensiunea de 48x48, fiind clasificate în 7 clase: furie, dezgust, frică, fericire, tristețe și surprindere, plus o clasă "neutră", ce cuprinde atât imaginile cu persoane ce nu exprimă nicio emoție vizibilă, cât și acelea ce redau emoții ce nu se încadrează în niciuna dintre cele 6 clase menționate anterior.

#### 4.1.2 LittleVGG

Partea centrală a aplicației, constând în recunoașterea emoțiilor pe baza expresiilor faciale, am realizat-o prin două abordări: prima dintre acestea folosește un model pre-antrenat pe baza de date [1]. Autorii folosesc o rețea neuronală convoluțională bazată pe arhitectura VGG, pe care însă o modifică din mai multe puncte de vedere: între altele, se folosesc și kernele de dimensiuni diferite de 3x3 și nu folosesc straturi fully-connected propriu-zise (adică straturi ascunse; se folosesc doar un strat de flattening — ce ar putea fi privit ca un strat de input — și unul de softmax — echivalentul stratului de ieșire; se pierde astfel neliniaritatea) la capătul rețelei. Aceste modificări sunt făcute în încercarea de a reduce complexitatea foarte mare a rețelei VGG (138 de milioane de parametri de antrenat). Ideea de bază a acesteia se păstrează însă, anume o oarecare regularitate a structurii și o adâncime relativ mică față de alte tipuri de rețele convoluționale folosite astăzi (de exemplu ResNet). Acuratețea obținută de acest model pe baza de date [4] este de 56%, conform autorilor. Redăm, în figura 1, arhitectura exactă a acestei rețele, botezată de autori "LittleVGG".

### 4.2 Recunoașterea emoțiilor

#### 4.2.1 Direcții de îmbunătățire explorate

În căutarea unei îmbunătățiri a acestui procent, am încercat la rândul nostru să creionăm o arhitectură de rețea bazată pe VGG, fiind însă constrânși de același aspect ca autorii LittleVGG: lipsa puterii de calcul pentru a antrena modele foarte adânci sau complexe (cu multe straturi/mulți parametri). Noutățile introduse de noi sunt adăugarea a două straturi de convoluție cu kernel de di-

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_1 (Conv2D) | (None, 46, 46, 32) | 320 |
| conv2d_2 (Conv2D) | (None, 44, 44, 64) | 18496 |
| max_pooling2d_1 (MaxPooling2 | (None, 22, 22, 64) | 0 |
| conv2d_3 (Conv2D) | (None, 20, 20, 128) | 73856 |
| max_pooling2d_2 (MaxPooling2 | (None, 10, 10, 128) | 0 |
| conv2d_4 (Conv2D) | (None, 8, 8, 128) | 147584 |
| max_pooling2d_3 (MaxPooling2 | (None, 4, 4, 128) | 0 |
| conv2d_5 (Conv2D) | (None, 4, 4, 7) | 903 |
| conv2d_6 (Conv2D) | (None, 1, 1, 7) | 791 |
| flatten_1 (Flatten) | (None, 7) | 0 |
| activation_1 (Activation) | (None, 7) | 0 |

Total params: 241,950
Trainable params: 241,950
Non-trainable params: 0

Figure 1: Arhitectura LittleVGG

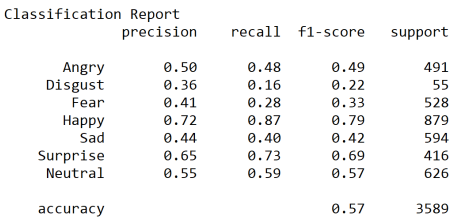
Model: "sequential_1"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv2d_1 (Conv2D) | (None, 48, 48, 32) | 320 |
| max_pooling2d_1 (MaxPooling2 | (None, 24, 24, 32) | 0 |
| conv2d_2 (Conv2D) | (None, 24, 24, 64) | 18496 |
| max_pooling2d_2 (MaxPooling2 | (None, 12, 12, 64) | 0 |
| conv2d_3 (Conv2D) | (None, 12, 12, 128) | 73856 |
| max_pooling2d_3 (MaxPooling2 | (None, 6, 6, 128) | 0 |
| conv2d_4 (Conv2D) | (None, 6, 6, 128) | 16512 |
| conv2d_5 (Conv2D) | (None, 6, 6, 128) | 16512 |
| flatten_1 (Flatten) | (None, 4608) | 0 |
| dense_1 (Dense) | (None, 7) | 32263 |

Total params: 157,959
Trainable params: 157,959
Non-trainable params: 0

Figure 2: Arhitectura CNN

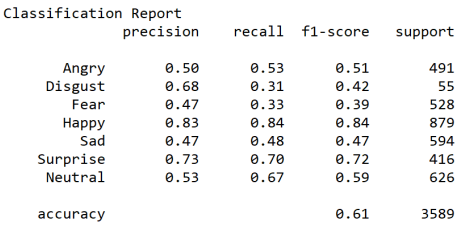
mensiune 1x1 și 128 de filtre, având ca scop creșterea non-liniarității (această decizie vine în compensarea lipsei straturilor fully-connected), precum și a unui zero-padding. Numărul total de parametri este aproximativ jumătate din cel al rețelei LittleVGG, iar acuratețea crește ușor și ea: de la 56 la 57 de procente. Acuratețea "globală" nu este însă cel mai exact mijloc de a măsura performanța modelului în acest caz, din cauză că setul de date nu conține un volum egal de exemple din fiecare clasă. Astfel, așa cum reiese și din figura 3, pentru unele clase — mai bine reprezentate în setul de antrenament — f1-score (media armonică dintre precizie și recall) se apropie de 80%, pe când în alte cazuri este sub 25%. Rezultatul este în mare măsură datorat faptului că această din urmă clasă (pe care modelul nu a reușit să o învețe) are de aproximativ 16 ori mai puține imagini decât clasa "happy" (cea care obține f1-score cel mai mare). Având în vedere sistemul de calcul folosit pentru antrenare (CPU și GPU de laptop, nu sisteme



Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Angry | 0.50 | 0.48 | 0.49 | 491 |
| Disgust | 0.36 | 0.16 | 0.22 | 55 |
| Fear | 0.41 | 0.28 | 0.33 | 528 |
| Happy | 0.72 | 0.87 | 0.79 | 879 |
| Sad | 0.44 | 0.40 | 0.42 | 594 |
| Surprise | 0.65 | 0.73 | 0.69 | 416 |
| Neutral | 0.55 | 0.59 | 0.57 | 626 |
| accuracy | | | 0.57 | 3589 |

Figure 3: Rezultatele testării propriei variante a VGG



Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Angry | 0.50 | 0.53 | 0.51 | 491 |
| Disgust | 0.68 | 0.31 | 0.42 | 55 |
| Fear | 0.47 | 0.33 | 0.39 | 528 |
| Happy | 0.83 | 0.84 | 0.84 | 879 |
| Sad | 0.47 | 0.48 | 0.47 | 594 |
| Surprise | 0.73 | 0.70 | 0.72 | 416 |
| Neutral | 0.53 | 0.67 | 0.59 | 626 |
| accuracy | | | 0.61 | 3589 |

Figure 4: Rezultatele testării ResNet-50

de tip cluster) și timpul de antrenare (doar aproximativ 3 ore, față de zile în cazul rețelelor mai complexe), considerăm acest rezultat ca o îmbunătățire față de rețeaua LittleVGG pe care am folosit-o ca punct de pornire. În figura 2, prezentăm arhitectura rețelei folosite de noi pentru generarea modelului.

Considerând performanțele obținute (departe totuși de cele pe care le întâlnim în prezent în cazul arhitecturilor CNN performante), precum și faptul că de la apariția arhitecturii VGG și până astăzi au mai fost propuse și alte abordări - cu rezultate îmbunătățite - ne-am stabilit ca obiectiv în continuare să folosim una dintre aceste variante de rețele convoluționale așa-zis "moderne" în încercarea noastră de a obține rezultate mai bune. Ca urmare, ne-am îndreptat atenția spre rețeaua ResNet, care este considerată în prezent un vârf de lance în domeniul clasificării automate de imagini. Din multitudinea de variante care se bazează pe ideea originală de "conexiuni reziduale" (*skip connections*), am ales ResNet-50 (are doar 50 de straturi în loc de 152 ale rețelei originale). Rezultatele au fost relativ mai bune — în sensul că nu mai există clase "neînvățate": cel mai mic f1-score este aproape 40% —, însă prețul pentru cele 4 procente în plus la acuratețe este un model semnificativ mai complex, pe lângă timpul de antrenare și de reglare a hiperparametrilor (learning rate și decay în principal) și el sensibil mai mare. Prezentăm în figura 4 rezultatele obținute de acest model pe fiecare clasă în parte.

# 5 Concluzii și îmbunătățiri posibile

Acest raport prezintă aplicarea a două arhitecturi de rețea neuronală convoluțională (VGG și ResNet) pentru recunoașterea de emoții pe baza informațiilor faciale,

observând că ResNet se descurcă mai bine (acurateţe cu 4% mai mare). Dintre direcţiile de îmbunătăţire posibile, menţionăm antrenarea pe un set de date mai mare şi/sau de calitate mai bună decât FER 2013, precum şi folosirea sistemelor de calcul de înaltă performanţă pentru antrenarea modelelor.

# References


[1] Facial Emotion Recognition (FER) Challenge - 2013 Dataset
https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

[2] Karolinska Directed Emotional Faces (KDEF) Dataset
http://kdef.se/

[3] Funcţionarea instrumentului pentru recunoaştere facială
https://medium.com/@ageitgey/machine-learning-is-fun-part-4-modern-face-recognition-with-deep-learning-c3cffc121d78

[4] Modele/tool-uri pentru detecţie de emoţii
https://github.com/priya-dwivedi/face_and_emotion_detection

[5] Model pentru recunoaştere facială
https://pypi.org/project/face_recognition/