

**UNIVERSITATEA DIN BUCUREȘTI FACULTATEA DE  
MATEMATICĂ ȘI INFORMATICĂ**

**LUCRARE DE LICENȚĂ**

**Coordonator:**  
**Prof. Dr. Radu Ionescu**

**Absolvent:**  
**Moldovan George-Alexandru**

București  
Iunie (fingers crossed), 2020

**UNIVERSITATEA DIN BUCUREȘTI FACULTATEA DE  
MATEMATICĂ ȘI INFORMATICĂ**

Sistem pentru detectarea anomaliilor in video

**Coordonator:**  
**Prof. Dr. Radu Ionescu**

**Absolvent:**  
**Moldovan George-Alexandru**

București  
Iunie (fingers crossed), 2020

# Cuprins

<b>1</b>	<b>Introducere</b>	<b>1</b>
1.1	Motivatie . . . . .	1
1.2	Context . . . . .	2
1.3	Conținutul lucrării . . . . .	3
<b>2</b>	<b>Analiza arhitecturii si a tehnologiilor folosite</b>	<b>4</b>
2.1	Etapa de antrenare . . . . .	4
2.2	Etapa de inferență . . . . .	6
2.3	Etapa de deployment . . . . .	6
<b>3</b>	<b>Sistemul de detecție a anomaliilor</b>	<b>9</b>
3.1	Detecția obiectelor . . . . .	9
3.2	Analiza antrenării sistemului . . . . .	10
3.3	Analiza etapei de inferență . . . . .	14
3.4	Scenarii de utilizare . . . . .	16
<b>4</b>	<b>Execuția în cloud a sistemului</b>	<b>18</b>
4.1	Cloud-computing în inteligență artificială . . . . .	18
4.2	Analiza opțiunilor . . . . .	19
4.3	Comparație între soluțiile PaaS si FaaS . . . . .	20
4.4	Descrierea soluției curente . . . . .	22

## Rezumat

Având în vedere contextul actual, detectarea anomaliilor în video este un subiect de interes în mai multe arii, în mod special în securitatea publică. Putem spune că această problemă este încă nerezolvată, deoarece sistemele actuale, deocamdată, nu depășesc omul cand vine vorba de detectarea anomaliilor. De asemenea, o altă problemă a sistemelor de detectare a anomaliilor în video este nevoia acestora de resurse computaționale mari în partea de inferență, făcând aproape imposibilă rularea acestora direct pe hardware-ul existent al sistemelor de supraveghere video actuale, acolo unde acestea prezintă un maxim interes. Astfel, putem spune că dezvoltarea unui sistem capabil să transforme sistemele de supraveghere actuale în sisteme ce pot recunoaște evenimente anormale este un subiect ce poate revoluționa domeniul supravegherii video. Această lucrare își propune o implementare al sistemului state-of-the-art la momentul redactării, așa cum este prezentat de *Ionescu et al.* [10]. Obiectivul este obținerea unei arhitecturi ce folosește o soluție PaaS si expunerea etapei de inferență printr-un API astfel încât convertirea unui sistem de supraveghere clasic într-unul inteligent să devină doar o problemă de implementare, fără a fi nevoie de schimbarea hardware-ului. Utilizarea unei soluții PaaS pentru etapa de inferență rezolvă problema executării cererilor fără complexitatea creeri și întreținerii unei infrastructuri de mașini virtuale sau fizice.

# Capitolul 1

## Introducere

### 1.1 Motivatie

Detectarea anomaliilor în video este în strânsă legătură cu sistemele de supraveghere inteligente, un domeniu care a fost și este de interes pentru mine. La rândul lor, sistemele de supraveghere inteligente, au o mare importanță în securitatea publică. Cu toții ne dorim o lume în care apelurile de urgență în caz de incendiu se fac automat, alunecările de teren sunt descoperite înainte să fie prea târziu, iar oamenii rău intenționați sunt opriți înainte să se întâmple tragedii.

Astfel, arhitectura folosită se bazează pe detecția caracteristicilor spațio-temporale ale evenimentelor prezente în video, care mai apoi sunt împărțite în clase de normalitate. Aceste caracteristici sunt extrase trecând evenimentul printr-o serie de auto-encodere pentru a folosi mai apoi reprezentarea latentă în cadrul clasificării finale. În etapa de antrenare, se folosesc filmări ce prezintă comportamentul normal în scenariul analizat. Analizând aceste video-uri putem crea un model capabil să recunoască dacă un eveniment aparține unei clase de normalitate analizate până acum, sau dacă este un caz anormal. Deoarece un eveniment poate fi ales în multe moduri, în cadrul acestui sistem un eveniment reprezintă orice obiect aflat în cadru. Analiza asupra fiecarui obiect conține și un cadru precedent dar și unul viitor, asemănător cu modalitatea folosită de *Ionescu et al.* [10]. O mențiune în acest sens ar fi că pentru fiecare obiect sunt analizate și cadrele de la pozițiile  $t-3$  și  $t+3$  respectiv la poziția  $t$  a cadrului analizat. Din acest motiv, atunci când sistemul va analiza un video în sistem live-feed, analiza se va face cu un decalaj de 3 cadre. Având în vedere că pentru un video actual viteza de redare este de minim 15 cadre pe secundă, acest decalaj este neglijabil.

Pe lângă partea algoritmică a detectării anomaliilor, o altă arie de interes a acestei lucrări este cloud computing. Această parte analizează un nou mod de rulare, ce facilitează atât dezvoltarea cât și execuția ulterioară a unor sisteme complexe. Acest nou mod constă în folosirea unei arhitecturi plasată în cloud, ce oferă dezvoltatorului posibilitatea să creeze sisteme ce necesită multe resurse în timpul rulării, fără costurile asociate creerii și menținerii unei infrastructuri proprii. Pe de altă parte, având în

vedere că toate operațiunile sunt executate în cloud, utilizatorii serviciului au nevoie doar de conexiune la internet și cerințe minime pentru sistemele proprii, fara a fi nevoiți să achiziționeze echipamente noi pentru a folosi sisteme de detecție a anomaliilor.

## 1.2 Context

Detectarea anomaliilor în video poate fi văzută ca o problemă subiectivă, deoarece un eveniment este normal sau anormal doar dacă este luat în considerare și contextul în care acesta apare. Un exemplu foarte bun este comparația între două persoane care se luptă și o persoană care se plimbă. Care dintre aceste evenimente este anormal ? Desigur, depinde de context. Dacă sistemul supraveghează o arenă de lupte, atunci persoana care se plimbă în ring prezintă un comportament anormal, în timp ce luptătorii prezintă comportamentul așteptat. Din acest motiv majoritatea lucrărilor din domeniu [5, 10, 14]... abordează un mod de lucru bazat pe antrenarea folosind video-uri ce provin din aceeași locație cu cele de test. Tocmai din cauza dependenței de context, detectarea anomaliilor nu este o problemă ce poate fi generalizată, astfel fiecare scenariu necesită o antrenare și un model propriu.

Ca și moduri de expunere a soluției software către utilizatori, aceasta se poate face în 2 moduri :

- Folosind servere proprii
- Folosind servicii cloud

Folosirea serverelor proprii presupune, pe lângă prezența fizică a serverelor și cumpărarea tuturor serviciilor conexe, cum ar fi servere de baze de date, sisteme de distribuire a fișierelor, infrastructură de rețea, ș.a.m.d., este nevoie și de o echipă dedicată pentru întreținerea infrastructurii și repararea eventualelor probleme ce pot apărea. Aceste considerente, împreună cu faptul că scalarea soluțiilor software este foarte anevoioasă atunci când este folosită infrastructura proprie, fac această soluție să nu mai fie folosită în mod curent deoarece încetinește dezvoltarea aplicației, iar rezultatul final este și el unul mai puțin calitativ decât ce se poate obține folosind soluții în cloud.

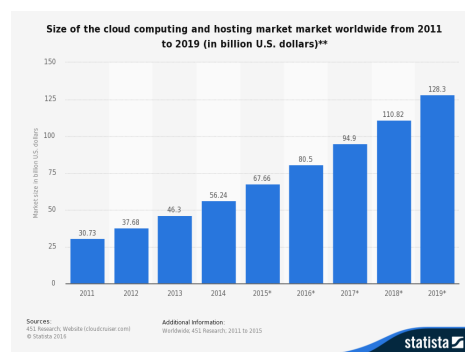


Figura 1.1: Statistică ce evidențiază importanța domeniului cloud computing în ultimii ani

Sursă: <https://www.statista.com/>

Folosirea serviciilor cloud oferă multiple posibilități de dezvoltare a aplicațiilor, de la crearea unei infrastructuri complete care este administrată în totalitate de către dezvoltator, la medii de execuție serverless care sunt complet administrate de către provider-ul de servicii cloud. Deși crearea unei infrastructuri proprii încă este necesară acolo unde legislația nu permite ca datele să fie stocate în cloud, în toate celelalte cazuri se face migrarea spre soluții cloud, lucru vizibil și în evoluția cotei de piață a domeniului cloud-computing la nivel global, așa cum se poate observa și în figura 1.1.

## 1.3 Conținutul lucrării

Ca și structură, lucrarea este împărțită în 2 părți:

- Partea algoritmică a sistemului de detectare a anomaliilor în video
- Partea de deployment a sistemului

În prima parte sistemul va fi analizat și detaliat din punct de vedere algoritmic și teoretic studiind problema detecției propriu zisă. Ca și tehnologii, în această parte am ales să folosesc Python3 ca și limbaj de programare pentru avantajele pe care le are. Printre acestea se numără faptul că este un limbaj orientat pe obiecte care pune la dispoziția dezvoltatorului numeroase librării specifice pentru AI/ML mutând astfel atenția dinspre detalii de implementare spre detalii de arhitectură și probleme mai abstracte ale programului care sunt cu adevărat importante pentru rezultatul final. Una dintre librăriile care s-a dovedit esențială dezvoltării este *Keras* [6], care oferă un API ce ușurează dezvoltarea unei rețele neuronale adânci/convoluționale dar și optimizează timpul de antrenare pentru modelele create.

În cea de a doua parte este analizat tipul de deployment al aplicației. Aici vor fi prezentate analize detaliate ale avantajelor și dezavantajelor soluțiilor cloud, dar și motivele pentru care modul final de deployment a fost ales.

Soluțiile de cloud-computing analizate vor fi:

- Infrastructure as a Service (IaaS) - analiză pe Amazon EC2
- Platform as a Service (PaaS) - analiză pe Amazon Elastic Beanstalk
- Function as a Service (FaaS) - analiză pe Amazon Lambda

# Capitolul 2

## Analiza arhitecturii si a tehnologiilor folosite

Pentru o mai bună înțelegere a contextului în care tehnologiile prezentate au fost folosite, acestea vor fi prezentate mai jos în cadrul secțiunii corespunzătoare locului în care a fost folosită în cadrul proiectului. Pentru toate etapele sistemului, limbajul folosit este Python3, deoarece, împreună cu librăriile și framework-urile existente pentru inteligență artificială, reprezintă mediul ideal pentru dezvoltarea unei soluții modulare, rapide și ușor de modificat.

### 2.1 Etapa de antrenare

Pentru etapa de antrenare, sistemul trece prin următoarea serie de etape:

- Detectează obiectele din setul de date
- Antrenează un autoencoder pe imaginile obiectelor și un alt autoencoder pe gradientii obiectelor.
- Obține reprezentarea latentă a fiecărui eveniment
- Stabilește  $k$  clase de normalitate folosind reprezentările latente
- Antrenează  $k$  clasificatori de tipul one-versus-rest

Ca prim pas, pentru obținerea obiectelor dintr-o imagine, atât în etapa de antrenare cât și în cea de inferență este folosit un detector de obiecte pre-antrenat. Un detector de obiecte este un sistem ce primește ca input o imagine și după procesarea acesteia rezolvă problema detecției de obiecte și întoarce pozițiile la care sunt plasate obiectele în imagine, și etichetele asociate acestora. Deși detectorul de obiecte face parte din arhitectură, deoarece detectarea de obiecte este un subiect în sine, implementarea unui astfel de algoritm nu face obiectul acestei lucrări. Datorită acestui lucru, este folosit un detector deja antrenat și testat pe setul de date COCO.



După aceasta etapă sunt antrenate cele două autoencodere convoluționale. Unul pentru imaginea propriu-zisă și altul pentru gradient. Un autoencoder este un tip de rețea neuronală alcătuit din 2 părți (encoder și decoder) care este folosit pentru a obține reprezentarea latentă a unui obiect folosind un mod de învățare nesupervizată. În timpul antrenării, autoencoder-ele au ca scop modificarea parametrilor interni pentru a obține la ieșire, datele primite la intrare. Deși ieșirea unui autoencoder nu prezintă interes, ceea ce este folositor este reprezentarea latentă (rezultatul encoder-ului) a datelor. Folosind această tehnică, se obține o reducere a dimensionalității ce îmbunătățește semnificativ performanțele clasificatorilor ulteriori.

Obținerea reprezentării latente a fiecărui obiect se realizează trecând imaginea, respectiv gradientul său prin encoderul autoencoderului corespunzător și păstrarea rezultatului.

Odată obținut vectorul de caracteristici pentru toate obiectele din setul de date, urmează stabilirea claselor de normalitate. Acest lucru se realizează prin aplicarea algoritmului k-means de clustering. Pentru implementare a fost folosit algoritmul *LLoyd* implementat în biblioteca python *sklearn*. Aplicând acest algoritm peste vectorii de caracteristici obținuți, rezultă k categorii de normalitate cu vectorii de caracteristici aferenți.

Folosind categoriile de normalitate obținute la pasul anterior, putem spune că față de o anumită categorie  $i$ , celelalte  $k-1$  categorii reprezintă categorii *artificial* anormale. Le numim *artificial* anormale deoarece în mod obiectiv ele sunt acțiuni normale pentru sistemul de detectare a anomaliilor, dar pentru antrenarea unor clasificatori conform schemei one-vs-rest acestea sunt tratate drept anormale. Astfel, putem antrena un clasificator binar  $g(i)$  în așa fel încât să separăm elementele din categoria  $i$  de cele din categoriile  $\{1,2,...k\}/i$  generând funcția :

$$f_i(x) = \sum_1^n w_j * x_j + b$$

. Unde  $x \in R^n$  reprezintă vectorul de caracteristici,  $w$  este vectorul de parametri a funcției, iar  $b$  reprezintă bias-ul funcției. [10].

Astfel, generăm k astfel de funcții corespunzătoare celor k clasificatori ce vor fi folosiți pentru a stabili dacă un eveniment este anormal. Conform schemei one-vs-rest, un eveniment este anormal dacă este clasificat drept anormal de către toți cei k clasificatori.

## 2.2 Etapa de inferență

În cadrul etapei de inferență, sistemul folosește detectorul de obiecte, autoencoderele pre-antrenate și cei  $k$  clasificatori binari pentru a stabili dacă un anumit eveniment este sau nu anormal. Astfel, parcursul sistemului este următorul :

- Extragerea cadrelor necesare din video
- Extragerea obiectelor din imagini
- Obținerea reprezentării latente
- Clasificarea evenimentelor

Pentru analiza unui eveniment care apare la un indice dat  $t$  sunt necesare 3 cadre. Mai precis cadrele de la indicii  $t-3$ ,  $t+3$  și  $t$ . Din cadrul  $t$  se va extrage vectorul de caracteristici specific aparenței vizuale, iar din celelalte 2 cadre se vor extrage vectorii de caracteristici specifici mișcării obiectului, prin analiza gradientilor. Prin concatenarea acestor 3 vectori, se obține vectorul final de caracteristici ce va fi folosit drept input pentru clasificatorii finali.

Pentru extragerea obiectelor din cadrele analizate, se va folosi același detector de obiecte ca în etapa de antrenare. Acesta va fi rulat pe cadrul principal  $t$ , urmând apoi să se folosească coordonatele obiectelor de la cadrul  $t$ , și pentru cadrele  $t+3$  și  $t-3$  deoarece din cauza diferenței mici de indici, obiectele nu se pot mișca îndeajuns încât să fie necesară rularea pe toate cele 3 cadre.

Odată obținute obiectele din cadrul analizat, după obținerea gradientilor din cadrele  $t-3$  și  $t+3$ , se poate obține reprezentarea latentă a acestor informații. Reprezentarea latentă a imaginii obiectului constă în rezultatul generat de encoderul autoencoderului pentru imagini iar reprezentarea latentă a gradientilor constă în rezultatul generat de encoderul autoencoderului pentru gradienti.

Odată obținuți vectorii de caracteristici pentru reprezentarea vizuală și pentru reprezentarea mișcării obiectului, prin concatenarea lor se obține vectorul final, ce poate fi folosit drept input pentru clasificarea finală. Conform schemei *one-vs-rest* acest vector este clasificat de toți cei  $k$  clasificatori, iar rezultatul final este scorul maxim obținut în urma clasificării.

## 2.3 Etapa de deployment

Pentru a face sistemul public, acesta este lansat drept un API ce rulează etapa de inferență pe un server web plasat în cloud. Pentru dezvoltarea serverului am ales să folosesc framework-ul *Flask*. Flask este un micro-framework de python folosit pentru dezvoltarea soluțiilor web. Motivele pentru care acest framework a fost ales sunt :

- Acesta adaugă un overhead foarte mic aplicației, lucru esențial atunci când aplicație se dorește a fi plasată în cloud, din cauza limitărilor de memorie.
- Oferă un suport foarte bun pentru planificarea rutelor de intrare în aplicație, lucru foarte important atunci când se dorește dezvoltarea unui API.
- Este unul dintre framework-urile suportate de Amazon Elastic Beanstalk

Comparativ cu alte framework-uri, Flask este diferit deoarece nu impune linii clare dezvoltatorilor atunci când vine vorba de forma sau componentele aplicației ce urmează a fi dezvoltată. Astfel, dezvoltatorul are control complet asupra aplicației și își poate manifesta creativitatea sau ideile fără a fi restricționat de framework. Flask a fost creat tocmai cu ideea de a fi construit peste el. Deși poate nu oferă aceeași viteză de dezvoltare comparativ cu celelalte frameworkuri, acesta oferă libertatea de alegere la fiecare pas. Are suport pentru toate tipurile de baze de date, fie ele relaționale sau nerelaționale, nu are preferințe când vine vorba de metode de autentificare sau de creare a rolurilor, totul este suportat și totul este la latitudinea dezvoltatorului. [9]

Pentru a lansa serverul în cloud, am folosit serviciul Amazon Elastic Beanstalk. Acesta este un serviciu complex, ce însumează la rândul lui mai multe servicii cloud oferite de Amazon Web Services(AWS). Mai jos sunt prezentate câteva dintre avantajele și dezavantajele acestui serviciu, fiind analizat în detaliu în capitolele ce urmează. Elastic Beanstalk este un serviciu de tipul *PaaS* ce oferă servicii de deployment și administrare complete. Pentru o mai bună detaliere, mai jos sunt definite toate serviciile cloud incluse de Elastic Beanstalk:

- Amazon EC2 : este un serviciu web oferit de Amazon ce constă în oferirea unui mediu de execuție sigur în cloud. Este echivalentul unei mașini fizice mutate în cloud. Acesta a fost creat pentru a ușura misiunea dezvoltatorilor de a migra serviciile proprii spre cloud computing. Aceste instanțe sunt extrem de configurabile, punând la dispoziția dezvoltatorului mai mult de 50 de tipuri de instanțe, plus diferite opțiuni de optimizare a unor părți specifice, cum ar fi memoria sau placa video. [2]
- Amazon S3 : este un serviciu ce oferă medii de stocare în cloud. Stocarea este de tipul cheie-obiect, unde cheia identifică unic la nivel global un fișier. Acesta oferă o securitate sporită a datelor, și o disponibilitate de 99.999999% deoarece datele sunt distribuite în sisteme diferite și în zone diferite. [4]
- Auto Scaling Group : acest serviciu oferă un mod automat de lansare a instanțelor EC2 astfel încât traficul să nu depășească niciodată puterea de execuție a unui aplicații. Scopul acestui serviciu este de a oferi capacitatea de a scala pentru a menține o performanță optimă, menținând costul în tot acest timp la valoarea minimă. [1]

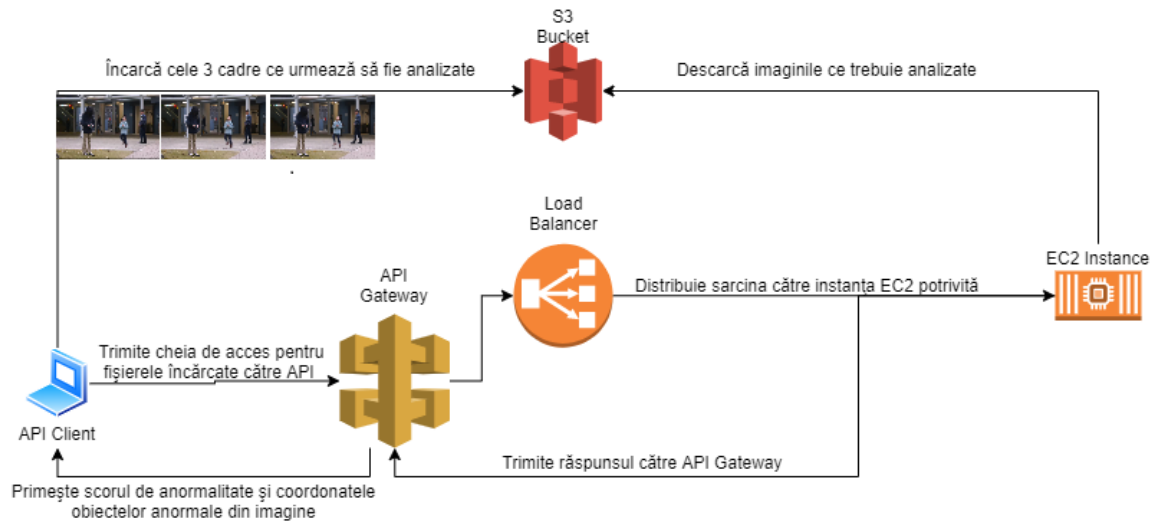


Figura 2.1: Arhitectura abstractizată a API-ului

- Elastic Load Balancing: este un serviciu care împarte traficul administrat de aplicație către instanțele lansate astfel încât acestea să fie utilizate într-un mod optim. Astfel, poate suporta încărcătura variabilă a aplicației și o poate distribui în așa fel încât, în funcție de modul ales (*accesibilitate crescută*, (*scalare automată*), *securitate ridicată*) aplicația să nu scadă sub performanțele dorite. [3]

Dezvoltatorul se ocupă doar de aplicația/serverul propriu zis, crează pachetul de deployment, iar Elastic Beanstalk crează instanțele EC2 necesare, administrează și rutează traficul către instanțe folosind un Load Balancer, iar atunci când aplicația este suprasolicitată, lansează în mod automat noi servere folosindu-se de avantajele unui Auto Scaling Group.

Așa cum se poate observa în figura 2.1, API-ul este creat în așa fel încât evaluarea se face pentru fiecare cadru în parte. Dezvoltatorul ce implementează clientul pentru API are doar responsabilitatea urcării cadrelor necesare într-un spațiu S3 prestabilit. Preprocesarea, extragerea obiectelor din imagine, și rularea detecției de anomalii sunt toate executate în cloud. Astfel, efortul computațional asupra clientului este minim. La momentul accesării API-ului, serverul trebuie să primească ca parametru în apelul HTTP cheia de acces pentru cadrele urcate de către client. Folosind această informație, pe server se descarcă aceste cadre și sunt analizate pentru a detecta anomaliile din cadrul central. Ca și rezultat, clientul primește scorul de anormalitate al cadrului împreună cu toate pozițiile obiectelor anormale din cadru, date ce pot fi folosite pentru notificări ulterioare sau diferite aplicații pe partea de client.

# Capitolul 3

## Sistemul de detecție a anomaliilor

În acest capitol va fi analizat în detaliu sistemul de detecție a anomaliilor în video, atât din punct de vedere al arhitecturii alese cât și din punct de vedere al implementării.

### 3.1 Detecția obiectelor

În ceea ce privește detecția de obiecte, aceasta este comună tuturor etapelor sistemului. Așa cum se poate observa și în secțiunea de analiză a performanțelor sistemului, detecția obiectelor din imagine joacă un rol esențial în viteza finală de analiză.

Prima opțiune a fost un detector de obiecte pre-antrenat pe setul de date COCO ce folosește o arhitectură **SSD-MobileNet**, din biblioteca *gluoncv model-zoo*. Acesta are ca prim avantaj viteza de analiză a unei imagini, dar și precizia bună a detecțiilor. Fiind asemănător cu cel folosit de *Ionescu et al.*, acesta folosește o arhitectură ce poate fi exploatată la potențialul ei maxim doar pe o mașină ce posedă un GPU puternic, având o viteza insuficientă pe sisteme bazate doar pe CPU.

Deoarece sistemul actual are ca scop final rularea pe o platformă ce necesită resurse minime, atât în cazul rulării offline cât și online, am ales folosirea unui model cu o arhitectură optimizată pentru rularea pe CPU. Mai precis, un model cu o arhitectură **YOLOv3-tiny** preluat din biblioteca *libcv*. Această schimbare a adus o îmbunătățire de **10x** în ceea ce privește timpul de detecție a obiectelor per cadru, de la o medie de 1200 de ms la 110 ms.

Detectorul YOLO este un tip de detector cu o singură etapă, ceea ce semnifică că nu mai este executată etapa de propunere a regiunilor întâlnite în detectoarele cu 2 etape, precum *R-CNN* sau *Fast R-CNN*. Astfel, imaginea este împărțită în  $S \times S$  pătrate. Fiecare pătrat este responsabil să detecteze obiectul aflat în interiorul acestuia. Fiecare pătrat are atribuit un scor ce arată probabilitatea ca un obiect să se afle în acesta, și încă un scor, ce reprezintă intersecția supra reuniune dintre perimetrul prezis și poziția adevărată a obiectului din imagine. Pentru fiecare astfel de pătrat sunt propuse *bounding box-uri* și scorul de încredere aferent. Un astfel de detector are la bază nivele convoluționale ce extrag caracteristici din imagine urmată de nivele conectate complet

pentru clasificare. Desigur, YOLOv3-tiny a suferit numeroase modificări în scopul optimizării arhitecturii din punct de vedere al timpului de execuție.[11] Odată cu viteza apare în schimb un dezavantaj în ceea ce privește detecția de obiecte: datorită arhitecturii minimaliste, detectorul ales detectează obiectele în integritatea lor doar dacă sunt relativ aproape de cameră și nu sunt foarte mici.

Implementarea detecției de obiecte a fost realizată la nivelul sistemului în cadrul clasei *ObjectDetector* ce prezintă următoarele funcționalități :

- La momentul instanțierii rulează detectorul de obiecte pre-antrenat pe imaginea primită ca parametru, obținând astfel marginile delimitatoare ale obiectelor din imagine împreună cu scorurile și clasele aferente.
- Oferă prin intermediul funcției *get\_object\_detections* obiectele extrase din imagine pe baza marginilor delimitatoare obținute în timpul instanțierii. Obiectele sunt redimensionate la  $64 \times 64$  și returnate sub forma unui vector de imagini.
- Oferă prin intermediul funcției *get\_detections\_and\_cropped\_sections* ce primește ca parametru și imaginea de la indexul  $t-3$  și imaginea de la indexul  $t + 3$  relativ la indexul  $t$  al imaginii folosite pentru instanțiere pachetul complet pentru trecerea prin sistemul de detecție și anume : Folosind marginile delimitatoare obținute în timpul instanțierii, sunt extrase obiectele din toate cele 3 imagini, și returnate sub forma a 3 vectori de imagini. Astfel un eveniment va fi format din cele 3 imagini aflate la un același index  $i$  în cei 3 vectori.

## 3.2 Analiza antrenării sistemului

Din cauza complexității problemei de a identifica comportamentul anormal al obiectelor prezente în video, arhitectura sistemului presupune multiple etape de prelucrare a datelor până la momentul clasificării finale. Astfel, etapa de antrenare este împărțită la rândul ei în 2 etape:

- Etapa de reducere a dimensionalității (antrenarea autoencoderelor)
- Etapa de antrenare a clasificatorilor finali

În etapa de reducere a dimensionalității sunt definite și antrenate autoencoderile pe baza imaginilor și gradientilor extrași din setul de date. Motivul pentru care obiectele sunt procesate de către autoencodere este că datorită antrenării doar pe obiectele din videourile de antrenare, acestea vor învăța să reprezinte doar evenimentele normale. Astfel, atunci când prin aceste autoencodere vor trece evenimente anormale, ce nu sunt asemănătoare cu cele de antrenament, autoencoderile vor genera o eroare de reconstrucție ce va ușura sarcina clasificatorilor finali.

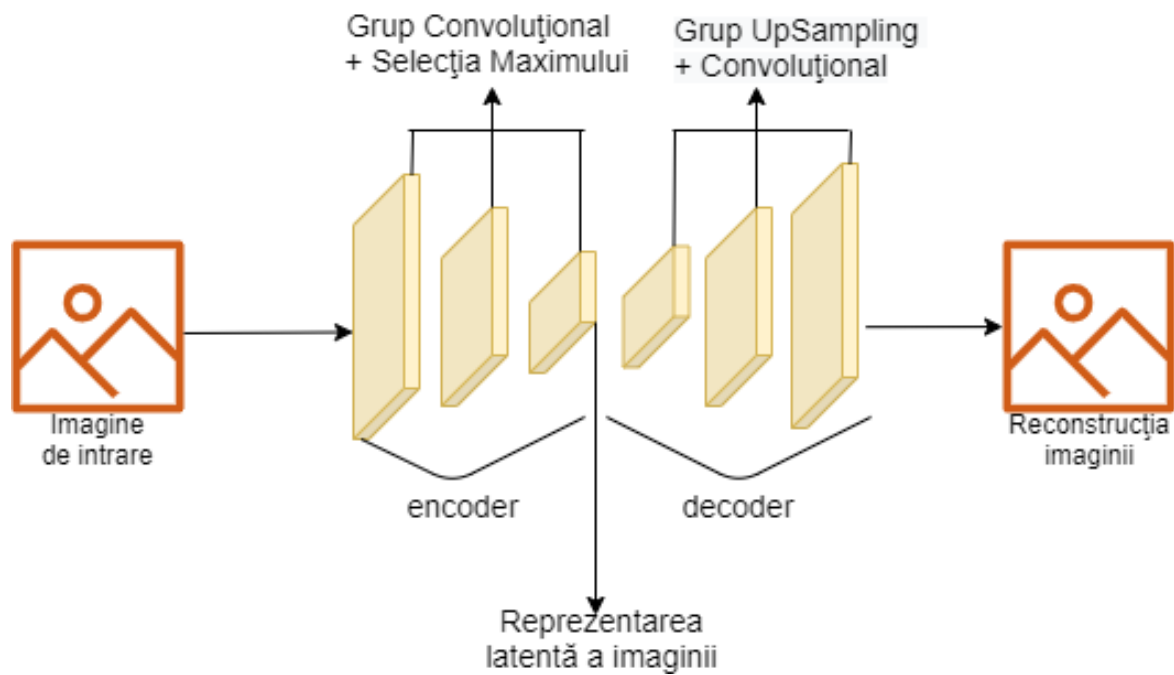


Figura 3.1: Prezentarea principiului de funcționare a autoencoderelor și arhitectura lor generală

Arhitectura aleasă pentru autoencodere este cea descrisă de *Ionescu et al.* [10] și constă într-o arhitectură convoluțională rapidă, formată dintr-un encoder cu 3 blocuri convoluțional + max-pooling și un decoder format din 3 blocuri upsampling + convoluțional. Fiecare autoencoder primește ca input date de dimensiune  $64 \times 64$  și creează după encoder un vector de caracteristici de dimensiune  $8 \times 8 \times 16$ . Structura detaliată a autoencoderelor este :

- Stratul de intrare, de dimensiune  $64 \times 64 \times 1$
- Bloc Convoluțional + MaxPooling format din : Un strat convoluțional bazat pe 32 de filtre de dimensiune  $3 \times 3$  urmat de funcția de activare *Relu* și un strat max-pooling bazat pe filtre  $2 \times 2$  cu *stride* 2.
- Bloc Convoluțional + MaxPooling format din : Un strat convoluțional bazat pe 32 de filtre de dimensiune  $3 \times 3$  urmat de funcția de activare *Relu* și un strat max-pooling bazat pe filtre  $2 \times 2$  cu *stride* 2.
- Bloc Convoluțional + MaxPooling format din : Un strat convoluțional bazat pe 16 de filtre de dimensiune  $3 \times 3$  urmat de funcția de activare *Relu* și un strat max-pooling bazat pe filtre  $2 \times 2$  cu *stride* 2.
- Bloc Convoluțional + UpSampling format din : Un strat convoluțional bazat pe 16 de filtre de dimensiune  $3 \times 3$  urmat de funcția de activare *Relu* și un strat UpSampling bazat pe filtre  $2 \times 2$ .

- Bloc Convoluțional + UpSampling format din : Un strat convoluțional bazat pe 32 de filtre de dimensiune  $3 \times 3$  urmat de funcția de activare *Relu* și un strat UpSampling bazat pe filtre  $2 \times 2$ .
- Bloc Convoluțional + UpSampling format din : Un strat convoluțional bazat pe 32 de filtre de dimensiune  $3 \times 3$  urmat de funcția de activare *Relu* și un strat UpSampling bazat pe filtre  $2 \times 2$ .
- Un ultim strat Convoluțional ce are ca rol reducerea dimensiunii de ieșire de la  $64 \times 64 \times 32$  la  $64 \times 64 \times 1[10]$  . Acesta este bazat pe 1 filtru de dimensiune  $3 \times 3$  urmat de funcția de activare *Sigmoid*

Primele 4 straturi reprezintă encoder-ul, în timp ce ultimele 4 straturi reprezintă decoderul.

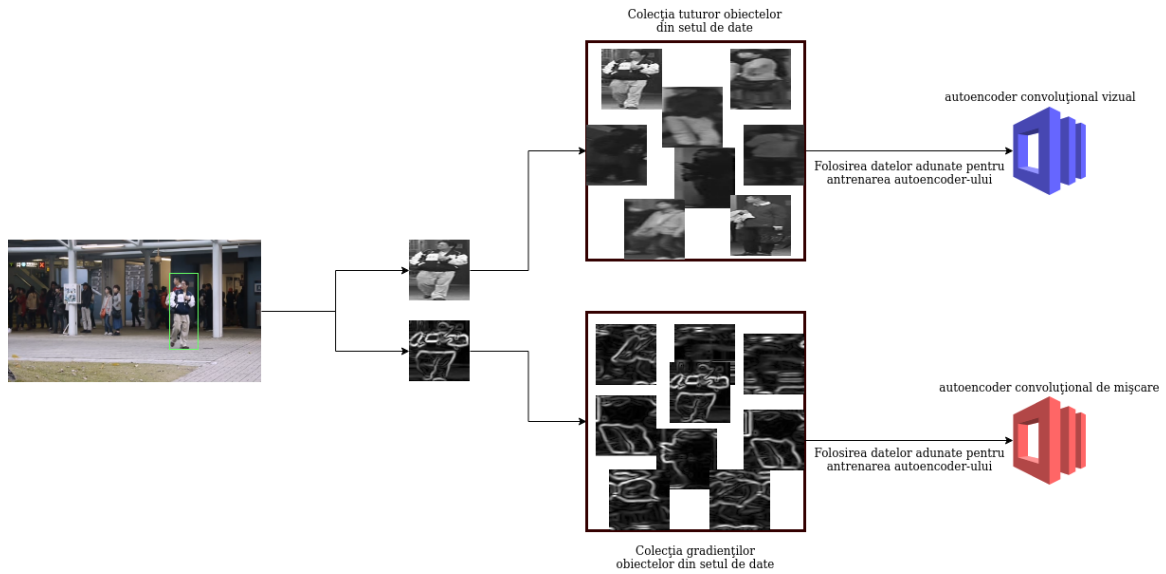


Figura 3.2: Arhitectura primei etape de antrenare

Implementarea arhitecturii prezentate în figura 3.2 constă în următorii pași:

- Pentru toate videourile de antrenare se execută detecția de obiecte pentru fiecare cadru, extrăgând astfel toate obiectele din video. Fiecare cadru este transformat în alb-negru pentru a fi prelucrat în etapele următoare. Pentru fiecare obiect extras, imaginea acestuia este redimensionată pentru a respecta dimensiunea de intrare a autoencodere-lor la  $64 \times 64$  iar apoi este calculat gradientul, ce reflectă mișcarea obiectului. Gradientul este calculat după formula:

$$\sqrt{G_x^2 + G_y^2}$$

Unde  $G_x$ ,  $G_y$  sunt imagini ce în fiecare punct conțin derivata orizontală respectiv verticală a imaginii inițiale, obținută prin aplicarea unui kernel Sobel de  $3 \times 3$ .



Dimensiunea kernelului folosit pentru calcularea gradientilor este foarte importantă pentru extragerea precisă a detaliilor, deoarece, așa cum se poate observa în figura 3.3, cu cât kernelul este mai mare, cu atât trasaturile extrase sunt mai generale.

- Imaginile și gradientii astfel obținuți sunt adăugați într-o colecție generală pentru tot setul de date. Cele 2 colecții astfel create vor servi drept input pentru antrenarea autoencoderelor.
- Folosind cele 2 colecții (de imagini, respectiv de gradienti) este antrenat câte un autoencoder pentru fiecare colecție. Înainte de a fi folosite pentru antrenarea autoencoderelor, atât imaginile, cât și gradientii, sunt normalizate în intervalul  $[0,1]$ . Antrenarea se realizează folosind optimizatorul Adam [13] și funcția loss ce constă în diferența medie a pătratelor, dată de formula :  $L(I, O) = \frac{1}{h*w} \sum_1^h \sum_1^w (I_{ij} * O_{ij})^2$  [10] unde I și O reprezintă imaginea de input respectiv de output și h, w sunt dimensiunile imaginilor. Aceasta este executată timp de 100 de epoci cu o rată de învățare de  $10^{-3}$ , dar având și o regulă de oprire rapidă, ce înseamnă că dacă timp de 2 epoci funcția loss nu se îmbunătățește, antrenarea este finalizată.

În etapa de antrenare a clasificatorilor finali, se folosesc autoencoderele antrenate în etapa precedentă pentru a obține vectorii de caracteristici specifici fiecărui eveniment prezent în setul de date. Ca și prim pas, se detectează toate obiectele prezente în fiecare cadru din video-urile de antrenare, iar pentru fiecare obiect, se folosesc coordonatele acestuia pentru a decupa același obiect din cadrul  $t-3$  și  $t+3$ , respectiv la indexul  $t$  curent. Se calculează gradientii pentru imaginile decupate, astfel este reprezentată mișcarea obiectului față de cadrul curent, iar apoi se obțin vectorii caracteristici ai fiecărui gradient prin trecerea acestora prin autoencoderul corespunzător antrenat în etapa precedentă. Odată obținuți cei 3 vectori caracteristici specifici evenimentului, așa cum este ilustrat și în figura 3.4, concatenarea acestora este stocată într-o colecție globală ce reține datele pentru tot setul de date.

Aplicând algoritmul de *k-means clustering* se obțin k categorii de normalitate. Astfel, pentru fiecare eveniment din colecția globală este cunoscută categoria  $i$  din care face parte. Folosind aceste date, putem antrena cei k clasificatori binari. Pentru fiecare categorie, un clasificator binar este antrenat folosind evenimentele din categoria curent drept date de antrenare pozitive, iar celelalte k-1 categorii drept date de

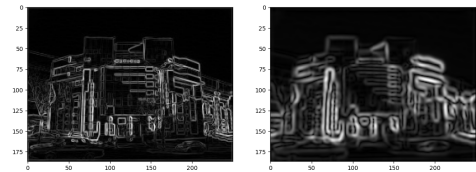


Figura 3.3: Gradienti obținuți în urma aplicării unui kernel  $3 \times 3$ , respectiv unul de dimensiune  $21 \times 21$

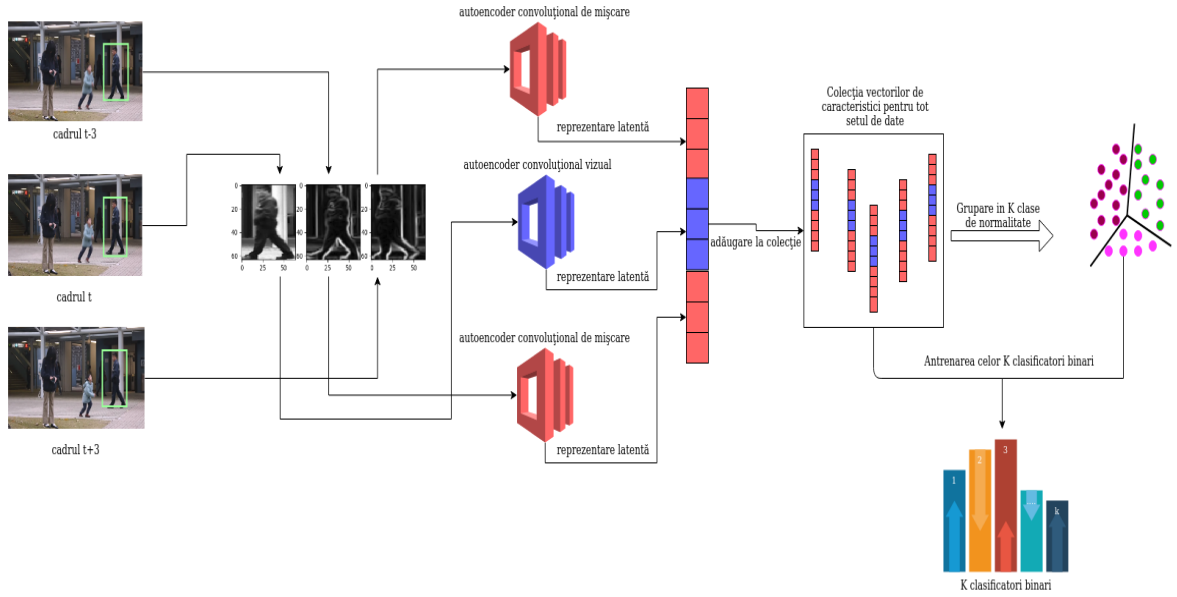


Figura 3.4: Arhitectura etapei finale de antrenare

antrenare negativă. În final, se obțin  $k$  clasificatori binari ce au rolul de a evalua dacă un eveniment aparține sau nu categoriei de normalitate  $i$  unde  $i$  este indicele clasificatorului  $g_i$ .

### 3.3 Analiza etapei de inferență

În cadrul etapei de inferență este analizat un eveniment cu scopul de a determina dacă acesta este sau nu anormal. Analiza unui singur eveniment stă la baza tuturor modurilor de utilizare a sistemului, deoarece, având aceste date, putem obține scoruri de normalitate pentru fiecare cadru, sau scoruri de normalitate la nivel de pixel sau chiar și clasificări generale de normalitate a unui video în totalitate.

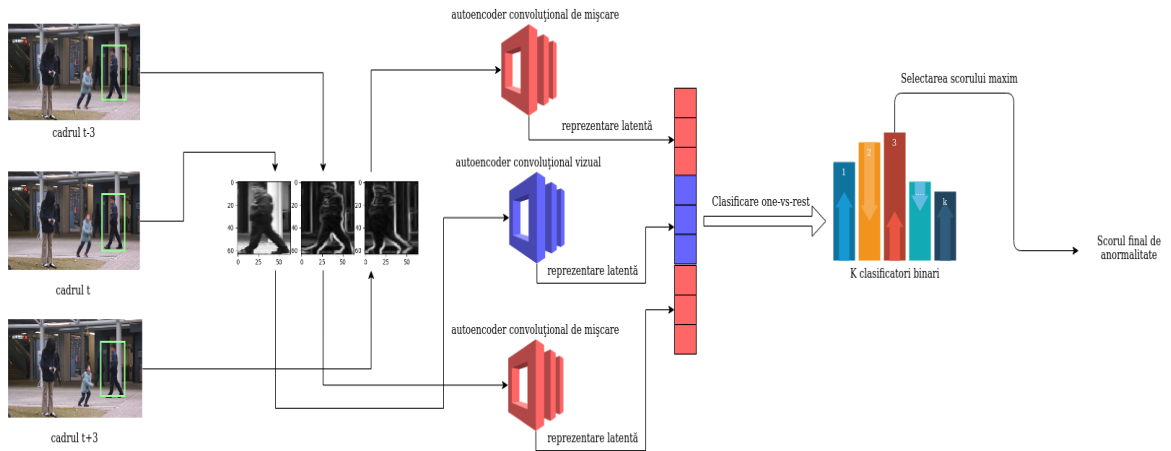


Figura 3.5: Arhitectura etapei de inferență

Pentru a obține scorul unui eveniment, un pas foarte important este obținerea vectorului de caracteristici corespunzător. Acesta se obține folosind autoencoderele antrenate în etapa precedentă, după pre-procesarea cadrelor corespunzătoare evenimentului. Pre-procesarea este asemănătoare etapei de antrenare, și anume, pentru fiecare obiect dintr-un cadru  $t$ , se selectează aceeași zonă din cadrele  $t-3$  și  $t+3$  cu scopul de a calcula gradientii ce reflectă mișcarea obiectului relativ la cadrul  $t$ . Odată obținute aceste 3 date (imaginea obiectului și cei 2 gradienti) așa cum este descris și în figura 3.5, acestea sunt trecute prin autoencodere și sunt obținute reprezentările latente, ce sunt concatenate pentru a se obține vectorul final de caracteristici. Vectorul de caracteristici este apoi clasificat de toți cei  $k$  clasificatori binari antrenați anterior, iar scorul final de anormalitate este dat de formula :

$$score = \max(g_i(v)), i \in [1..k]$$

unde  $v$  este vectorul de caracteristici, iar  $g_i$  este un clasificator binar.

Filmarea propriu zisă și selecția cadrelor este realizată pe partea de client, în timp ce pre-procesarea cadrelor, detecția obiectelor și calcularea scorului pentru fiecare dintre aceste obiecte este executat pe server. Astfel, atât timp cât există o conexiune la rețea, clientul nu este limitat de puterea de calcul proprie, și poate analiza mai multe videoclipuri simultan. Totuși clientul trebuie să încarce cadrele ce vor fi analizate către destinația S3 a serverului. Pentru a nu încetini semnificativ procesul de analiză, trimitere cadrelor în rețea trebuie făcută în paralel, folosind eventualele posibilități de multi-threading ale clientului.

În ceea ce privește serverul pe care rulează analiza propriu zisă, acesta este un HTTP API, ce expune pentru public capacitatea de a rula inferența. În stadiul curent, API-ul este configurat cu un singur endpoint, de tip POST : “upload/frame\_key” , unde `frame_key` reprezintă cheia de acces cu care au fost urcate în cloud cadrele ce urmează să fie analizate. Pentru ca o cheie de acces să fie considerată validă de către server, acesta trebuie să găsească în destinația S3 specifică 3 fișiere ce au următoarele chei :

- `frame_key`
- `frame_key_d3`
- `frame_key_p3`

`Frame_key` este parametrul primit prin HTTP, iar `frame_key_d3` și `frame_key_p3` sunt chei obținute prin concatenarea sufixului `_d3` respectiv `_p3` la cheia primită ca parametru. Dacă parametrul primit reprezintă o cheie validă, atunci cadrele sunt descărcate temporar pe server, unde sunt folosite pentru a rula întregul proces de inferență, iar în final, rezultatele sunt transmise clientului în format json. JSON-ul rezultat conține 3 câmpuri :

- Codul de stare: ce reprezintă codul http rezultat în urma apelului

- Result: ce conține scorul cadrului de anormalitate, ce reprezintă maximum dintre scorurile obiectelor din cadru.
- boxes: o listă ce conține coordonatele obiectelor ce sunt considerate anormale din cadrul principal

Codul de stare respectă standardul HTTP1.1, mai precis *RFC7231* [8] și respectă valorile prezentate în tabelul de mai jos:

Valoarea codului	Semnificație
1xx	Apelul a fost primit, dar procesarea încă este în desfășurare
2xx	Apelul a fost primit și procesat cu succes
3xx	Clientul trebuie să execute pași suplimentari pentru ca apelul să fie procesat
4xx	Apelul este greșit din punct de vedere sintactic sau nu poate fi procesat
5xx	Apelul este valid dar serverul a întâmpinat o eroare în timpul procesării

Tabela 3.1: Descrierea codurilor returnate de către server

Modul în care este structurat serverul face foarte ușoară crearea de noi API-uri, cu aceeași funcționalitate, dar orientate spre tipuri de anormalitate diferite. Așa cum a fost descris și în introducere, detectarea anomaliilor este dependentă de setul de date de referință. Astfel, fiecare model antrenat pe un set de date diferit, acoperă o arie diferită de anomalii. Altfel spus, în funcție de setul de date folosit la antrenare, sistemul va detecta un alt set de evenimente drept anomalii. În fine, un nou API este necesar pentru fiecare astfel de model.

Deoarece arhitectura curentă a serverului încarcă modelele pre-antrenate dintr-o destinație S3 la momentul inițializării, pentru a crea un nou API ce a fost antrenat pe un alt set de date, este necesară doar crearea unei noi destinații S3, încărcarea modelelor pre-antrenate la acea destinație, schimbarea în cod a denumirii vechii destinații în cea nouă și apoi un nou deployment. Astfel, cu un număr minim de pași, poate fi creat un nou API ce execută etapa de inferență pentru un sistem ce a fost antrenat pe un nou set de date.

## 3.4 Scenarii de utilizare

Acest sistem de detecție a anomaliilor are aplicații multiple în domeniul supravegherii inteligente. Este cunoscut faptul ca datele înseamnă putere și control, însă acest lucru este adevărat doar atunci când datele sunt analizate. Există nenumărate sisteme de supraveghere actuale ce nu conțin nici o formă de analiză pasivă, sub forma unui program. Majoritatea se bazează pe monitorizarea activă a datelor de către o persoană, fapt ce limitează capacitatea sistemului de a fi monitorizat, deoarece pentru pentru

supravegherea unei suprafețe mari, nu este posibilă monitorizarea activă a tuturor datelor.

Deoarece sistemul actual a fost dezvoltat să ruleze folosind relativ puține resurse, acesta este pretabil pentru a fi folosit atât pe echipamente independente din industria IoT cât și pe sisteme complexe de analiză a datelor agregate din mai multe surse.

Un scenariu de utilizare în industria IoT, ar fi detecția braconajului, sau a defrișărilor de pădure. Deoarece de obicei conexiunea la internet în parcurile naturale este foarte slabă, trimiterea imaginilor către server nu este o opțiune. Astfel, sistemul trebuie rulat local, pe dispozitivul propriu-zis, urmând mai apoi, în cazul detectării unei anomalii, să fie trimisă către server doar o notificare, fapt ce reduce semnificativ cantitatea de date trimisă prin rețea. În acest caz, sistemul ar putea fi combinat cu un senzor fizic de mișcare, pentru a analiza doar imaginile în care există mișcare. Astfel, consumul de energie al dispozitivului ar fi de asemenea redus.

În ceea ce privește sistemele de analiză a datelor, sistemul actual este pretabil pentru a fi distribuit într-o rețea de servere și pentru a analiza în paralel înregistrări din surse multiple. În acest caz, sistemul poate analiza atât date în timp real, cât și date istorice în funcție de necesități. Astfel, sisteme complexe precum sistemul de supraveghere al unui oraș poate fi monitorizat pasiv, iar monitorizarea activă să se facă doar asupra anomaliilor detectate de sistem. Astfel, șansele ca anumite evenimente să treacă neobservate sunt reduse semnificativ.

# Capitolul 4

## Execuția în cloud a sistemului

### 4.1 Cloud-computing în inteligență artificială

Execuția în cloud este un domeniu cu o creștere substanțială în ultimii ani, înlocuind practic o bună parte din infrastructurile clasice existente. Acest lucru are implicații și în dezvoltarea sistemelor de inteligență artificială, acesta fiind un domeniu știut drept unul cu necesar de putere de execuție mare, dar și cu un potențial de scalare extraordinar. Tocmai acest potențial, face ca inteligența artificială să fie candidatul perfect pentru execuția în cloud. Faptul că efortul de dezvoltare pentru a crea un sistem în cloud pregătit să deservească milioane de utilizatori este egal cu dezvoltarea unui sistem pentru câteva mii de utilizatori în mod clasic, arată din nou de ce această soluție a devenit alegerea perfectă pentru multe companii.

Deși opțiunile pentru execuția în cloud sunt vaste, de la sisteme complet administrate de către utilizator, până la funcții cloud complet administrate de distribuitor, tendința este ca acolo unde este posibil, clientul să se ocupe cât mai puțin de administrare, și cât mai mult de dezvoltarea propriu zisă a produsului. Un alt subiect de interes pentru execuția în cloud este comparația între arhitecturi *serverfull* și *serverless* dar alegerea între cele două diferă de la sistem la sistem deoarece pentru a alege o arhitectură *serverless*, sistemul trebuie construit pentru asta încă de la începutul dezvoltării.

Sistemele de ML, sau IA, sunt de obicei folosite pentru a îmbunătăți sisteme deja existente, sau pentru a ușura luarea deciziilor vitale. Din acest motiv, monitorizarea execuției, asigurarea disponibilității sistemului și posibilitatea de a controla costurile sunt lucruri foarte importante pentru alegerea modului de execuție a sistemului. Pe lângă asigurarea tuturor acestor facilități, soluțiile cloud oferă de multe ori și performanțe mai bune decât infrastructurile fizice echivalente, oferind astfel toate bazele necesare pentru găzduirea sistemelor IA/ML.

La ora actuală, primii 3 distribuitori de soluții cloud (Amazon, Microsoft, Google) oferă și soluții speciale de implementare rapidă a inteligenței artificiale în aplicații deja existente, folosind sisteme special gândite să suporte întreg procesul de dezvoltare și execuție, totul înglobat într-un singur serviciu :

- Amazon SageMaker
- Google AI Platform
- Azure Machine Learning

## 4.2 Analiza opțiunilor

Opțiunile în ceea ce privește execuția în cloud sunt diverse și în continua dezvoltare. Deși categoriile de soluții sunt bine definite, serviciile propriu zise sunt modificate destul de des, având din ce în ce mai multe facilități.

Categoriile de servicii cloud sunt :

- Infrastructure as a Service (IaaS)
- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Function as a Service (FaaS)

Soluțiile de tip *IaaS* presupun punerea la dispoziția clientului doar a infrastructurii cerute, fara nici o administrare din partea distribuitorului. Din acest punct de vedere, soluțiile IaaS pot fi vazute de către client drept infrastructuri fizice, singura diferență fiind că locul în care este amplasată aparatura nu este deținută de către client. În ceea ce privește securitatea, update-urile, softurile și toate legăturile din cadrul infrastructurii, acestea sunt administrate strict de către client. Între a rula un sistem pe o infrastructură locală și a rula un sistem pe o infrastructura aflată în cloud dar folosind o soluție IaaS există doar diferențe de natură economica.

*SaaS* reprezintă o metoda pentru distribuirea softurilor pe internet la cerere și pe bază de subscripție. Soluțiile SaaS sunt folosite pentru găzduirea și administrarea softurilor pentru a facilita distribuția de update-uri sau patchuri de securitate.

Soluțiile *PaaS* reprezintă un mod de a crea medii de dezvoltare, testare sau producție la cerere. Este construit pentru a asigura o modalitate rapidă de a crea aplicații software diverse, de la aplicații web, la cele mobile sau API-uri ce fac parte dintr-un alt sistem. Practic, este un mod de a crea și administra soluții IaaS în mod automat, asigurând toate uneltele necesare pentru scalabilitate automată, distribuirea traficului între servere, mentenanță, asigurarea securității și lansarea de noi versiuni. Deși este un serviciu serverfull, acesta ia de pe umerii clientului sarcina creeri și

administrării infrastructurii, grăbind astfel procesul de dezvoltare și lansare a unui nou sistem software.

În ceea ce privește FaaS, acesta este un domeniu nou, deoarece a apărut pentru prima dată în 2010 fiind oferit de câteva start-upuri la acea vreme. Acest mod de dezvoltare orientat spre microservicii a devenit trendul în industrie în ultimii ani pentru sisteme cu potențial de scalare mare, deoarece prezintă numeroase avantaje din punct de vedere al modului de dezvoltare și de execuție în industrie. În momentul de față, pentru servicii de tip FaaS sunt 3 mari jucători: Amazon cu AWS Lambda, Google cu Google Cloud Functions și Microsoft cu Azure Functions.[12]. Numeroase lucrări din domeniu [7, 15] arată ca rularea algoritmilor de machine learning folosind soluții FaaS (Function as a service) precum AWS Lambda sau Google cloud functions, este în sine o problemă ce necesită soluții de optimizare a codului pentru a îndeplini restricțiile soluțiilor de rulare serverless, cum ar fi memoria limitată a mediului de execuție.

O altă caracteristică a soluțiilor cloud este tipul de execuție și anume: *serverfull* sau *serverless*. Prin *serverfull*, se înțelege o soluție ce rulează pe servere ce pot fi indentificate în mod unic, care rulează în mod continuu, dar pe care se execută diferite operații în funcție de nevoile sistemului. IaaS și PaaS sunt mereu soluții *serverfull*, în timp ce un serviciu SaaS poate fi atât *serverfull* cât și *serverless*. Prin *serverless* se înțelege o soluție ce lansează noi micro-instanțe pentru fiecare cerere, ce este mai apoi oprită după execuția codului. Costul unei soluții *serverless* este direct proporțional cu timpul de execuție, în timp ce pentru soluțiile *serverfull* costul este constant.

Caracteristicile celor 2 tipuri sunt prezentate în tabelul de mai jos :

Caracteristici	Soluții AWS <i>serverfull</i>	Soluții AWS <i>serverless</i>
Când este activ serviciul	Când este declanșat de un eveniment	Continuu, până la oprire
Limbaaj de programare	Python, Java, Go, C#, și altele...	Orice limbaj
Max RAM	0.125 - 3 Gb	0.5-1952 Gb
Max Capacitate de stocare	0.5 Gb	0-3600 Gb
Max Timp de rulare	900 secunde	Nelimitat
Unitatea minimă taxată	0.1 secunde	60 de secunde
Preț minim pe unitate	\$0.0000002	\$0.0000867
Sistem de operare	Ales de distribuitorul de soluții cloud	Ales de către client

### 4.3 Comparație între soluțiile PaaS și FaaS

Atât soluțiile PaaS cât și cele FaaS sunt pretabile sistemelor de inteligență artificială, deoarece oferă administrare automată, putere mare de calcul și avantajul că suportă optimizări prin rularea pe GPU.



Soluțiile PaaS, deși oferă servicii de administrare automată, în spatele acestui mecanism se ascunde tot o infrastructură clasică, cu servere fizice ce rulează în mod continuu, ce necesită echilibrarea traficului între acestea, lansarea de noi servere, lansarea de noi versiuni a aplicației pe serverele deja pornite sau închiderea unor servere atunci când nu mai sunt folosite. Toate acestea sunt realizate de către distribuitorul de soluții cloud, punând la dispoziția clientului un mediu gata să suporte orice tip de trafic către aplicație. Astfel dezvoltatorul se poate gândi doar la construirea unui server cât mai eficient, pentru a folosi resursele unei instanțe la maxim. Un aspect important pentru soluțiile PaaS este costul asociat instanțelor folosite. Deși numărul lor este variabil, în orice moment măcar o instanță trebuie să fie activă. Asta înseamnă că dacă sistemul poate rula doar pe instanțe cu resurse suplimentare, ce au un cost orar mai mare, sistemul va avea un cost de operare mare chiar și atunci când este subutilizat. Un sistem ce folosește soluții PaaS în mod ideal este de preferat să ruleze pe mai multe instanțe cu capacitate de execuție mică față de mai puține instanțe cu capacitate mare, deoarece lansarea sau închiderea de noi instanțe este gratuită, în timp ce rularea acestora atunci când sunt sub-utilizate nu este.

Soluțiile FaaS, pe de altă parte, abstractizează și mai mult infrastructura existentă, luând orice responsabilitate de pe umerii clienților atunci când vine vorba de administrarea acestora. FaaS presupune existența unei funcții, ce simbolizează o aplicație, ce este executată de fiecare dată când apare un eveniment declanșator. Avantajul acestei abordări este costul asociat, și anume faptul că plata se realizează strict pentru timpul în care funcția a fost executată. Timpul petrecut pentru a aștepta apariția unui eveniment nu este taxat. Spre exemplu, soluție FaaS a celor de la Amazon, AWS Lambda, poate fi declanșată în numeroase moduri, de la încărcarea unui fișier într-o destinație S3 prestabilită, la un apel HTTP către un endpoint atașat funcției. Aceste soluții suportă majoritatea limbajelor de programare, și pot fi configurate în așa fel încât să acopere orice nevoie în ceea ce privește puterea de execuție. Soluțiile FaaS prezintă și o serie de dezavantaje printre care se numără:

- Există o limită asupra dimensiunii codului ce urmează să fie executat de către funcție.
- Timpul de execuție a unei funcții este de asemenea limitat.
- Starea programului nu este păstrată între execuții făcând necesară stocarea datelor ce trebuie să persiste în locații dedicate (S3 sau o bază de date)
- Limitarea executării concurente. (Lambda nu permite mai mult de 1000 de funcții să fie executate concomitent)

Deși din descrierea acestui tip de soluții, pare alegerea ideală pentru execuția etapei de inferență pentru un model de *machine learning*, situația devine complicată atunci când sistemul este unul complex, iar respectarea acestei limite de dimensiune devine

un blocaj. De multe ori, aplicațiile de AI/ML depind de biblioteci și frameworkuri cunoscute în domeniu precum *tensorflow*, *keras*, *mxnet*, dar care au fost dezvoltate având în minte eficiența execuției, și nu minimizarea codului sursă. Având în vedere ca AWS Lambda limitează dimensiunea codului atașat la 250MB, lansarea sistemelor ce se bazează pe rețele neuronale este o sarcină grea din punct de vedere al încadrării în aceste restricții.

## 4.4 Descrierea soluției curente

Soluția aleasă pentru execuția etapei de inferență în cloud este *AWS Elastic Beanstalk*. Acest serviciu de tip PaaS oferă mediul perfect pentru lansarea aplicației, atât din punct de vedere al configurabilității cât și al securității oferite. Pentru a lansa o aplicație folosind elastic beanstalk, este necesar să creezi un pachet de lansare, ce conține codul sursă și celelalte instrucțiuni pentru inițializarea serviciului. Deși acesta este limitat la 512MB, este mai mult decât îndeajuns deoarece mediul de rulare și dependențele programului nu sunt adăugate în acest pachet, ci sunt descarcate pe server la momentul lansării pe baza unui fișier de configurare adăugat în pachet numit *requirements.txt*.

În timpul lansării aplicației, sunt create automat următoarele resurse :

- O flotă , sau o singură instanță *EC2*, în funcție de configurație
- Un *Elastic Load Balancer* ce distribuie traficul între instanțe
- Un *AutoScaling Group* ce definește regulile de scalare a numărului de instanțe
- O destinație *S3* în care este stocat codul sursă a aplicației
- Și un *API Gateway* ce joacă rolul de punct de intrare al traficului în aplicație

Deoarece la bază se află instanțe EC2, aceasta conferă un mare avantaj din punct de vedere al execuției concurente acestei soluții față de AWS Lambda sau alte soluții FaaS asemănătoare. Deși mai multe instanțe Lambda pot fi executate concurent, în cadrul unei singure instanțe, apelurile concurente nu sunt permise. Luând în calcul că în etapa de inferență sunt și părți de input/output în care resursele sunt nefolosite, folosirea instanțelor EC2 ce pot rula mai multe apeluri în mod concurent este opțiunea optimă din punct de vedere al folosirii resurselor.

Capacitatea de multi-threading din cadrul serverelor, la care se adaugă scalarea automată a numărului acestora în funcție de traficul ce accesează API-ul, face ca resursele să fie folosite la maxim, atât din punct de vedere al performanței disponibile, cât și din punct de vedere economic.

Un alt avantaj al acestei soluții este ușurința cu care se poate schimba mediul de execuție. Spre exemplu, într-un anumit scenariu, în care se observă în panoul de

monitorizare ca instanțele EC2 curente sunt suprasolicitate, cu ajutorul interfeței web, sau cea din linia de comandă, pot fi schimbate tipurile de instanțe EC2 folosite cu unele mai performante fără a avea serviciul indisponibil pe durata acestei modificări.

# Listă de figuri

1.1	Statistică ce evidențiază importanța domeniului cloud computing în ultimii ani . . . . .	2
2.1	Arhitectura abstractizată a API-ului . . . . .	8
3.1	Prezentarea principiului de funcționare a autoencoderelor și arhitectura lor generală . . . . .	11
3.2	Arhitectura primei etape de antrenare . . . . .	12
3.3	Gradienți obținuți în urma aplicării unui kernel $3 \times 3$ , respectiv unul de dimensiune $21 \times 21$ . . . . .	13
3.4	Arhitectura etapei finale de antrenare . . . . .	14
3.5	Arhitectura etapei de inferență . . . . .	14

# Listă de tabele

3.1	Descrierea codurilor returnate de către server . . . . .	16
-----	--	----

# Bibliografie

- [1] Amazon autoscaling description, 04 2020.
- [2] Amazon ec2 description, 04 2020.
- [3] Amazon elastic load balancing description, 04 2020.
- [4] Amazon s3 description, 04 2020.
- [5] K. Cheng, Y. Chen, and W. Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. pages 2909–2917, 2015.
- [6] F. Chollet et al. Keras description, 04 2020.
- [7] A. Christidis, R. Davies, and S. Moschoyiannis. Serving machine learning workloads in resource constrained environments: a serverless deployment example. pages 55–63, 2019.
- [8] R. Fielding and J. Reschke. Hypertext transfer protocol (http/1.1): Semantics and content. RFC 7231, RFC Editor, June 2014.
- [9] M. Grinberg. *Flask Web Development: Developing Web Applications with Python*. O’Reilly Media, Inc., 1st edition, 2014.
- [10] R. T. Ionescu, F. S. Khan, M. Georgescu, and L. Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. pages 7834–7843, 2019.
- [11] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu. A survey of deep learning-based object detection. *IEEE Access*, 7:128837–128868, 2019.
- [12] E. Jonas, J. Schleier-Smith, V. Sreekanti, C.-C. Tsai, A. Khandelwal, Q. Pu, V. Shankar, J. Carreira, K. Krauth, N. Yadwadkar, et al. Cloud programming simplified: A berkeley view on serverless computing. *arXiv preprint arXiv:1902.03383*, 2019.
- [13] Peng Qi, Wei Zhou, and Jizhong Han. *A method for stochastic L-BFGS optimization*. 2017.

- [14] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. pages 6479–6488, 2018.
- [15] H. Wang, D. Niu, and B. Li. Distributed machine learning with a serverless architecture. pages 1288–1296, 2019.