Universitatea “Politehnica ” din București

Facultatea de Electronică, Telecomunicații și Tehnologia Informației

**Descrierea conținutului vizual pe terminalul mobil**

**Lucrare de disertație**

**prezentată ca cerință parțială pentru obținerea titlului de**

**Master în domeniul Imagisticii Digitale**

**programul de studii de Tehnici Avansate in Imagistica Digitală**

Conducători științifici Absolvent

*Prof. dr. ing. Corneliu FLOREA George-Iulian NITROI*

*As. drd. ing Badea Mihai*

Anul 2020



Copyright © 2020 , George-Iulian Nitroi

Toate drepturile rezervate

Autorul acordă UPB dreptul de a reproduce și de a distribui public copii pe hârtie sau electronice ale acestei lucrări, in formă integrală sau parțială

|  |
| --- |
|  |

Declarație academica

**Cuprins**

[Listă de figuri 11](#_Toc42678354)

[Listă de Tabele 12](#_Toc42678355)

[Listă de acronime 14](#_Toc42678356)

[Introducere 16](#_Toc42678357)

[Capitol 1 Noțiuni teoretice necesare dezvoltării practice 18](#_Toc42678358)

[1.1 Topologia aplicației 18](#_Toc42678359)

[1.2 Terminalul mobil 18](#_Toc42678360)

[1.2.1 Platforma Android 19](#_Toc42678361)

[1.2.2 Hardware 19](#_Toc42678362)

[1.3 Firebase 19](#_Toc42678363)

[1.4 Python 20](#_Toc42678364)

[1.5 Flask 20](#_Toc42678365)

[1.6 Tensorflow 21](#_Toc42678366)

[1.6.1 Modul de funcționare 21](#_Toc42678367)

[1.6.2 Keras 21](#_Toc42678368)

[Capitol 2 Modelul Rețelei Neuronale 23](#_Toc42678369)

[2.1 Rețelele convoluționale 24](#_Toc42678370)

[2.2 Rețele recurente 29](#_Toc42678371)

[2.2.1 Gated recurrent unit 30](#_Toc42678372)

[2.2.2 Straturile de embedding 31](#_Toc42678373)

[2.3 Transfer learning 32](#_Toc42678374)

[2.4 Baza de date 33](#_Toc42678375)

[2.5 Implementare și Antrenare 33](#_Toc42678376)

[2.5.1 Antrenarea 38](#_Toc42678377)

[2.6 Rezultate 39](#_Toc42678378)

[Capitol 3 Aplicația Android 42](#_Toc42678379)

[3.1 Introducere 42](#_Toc42678380)

[3.2 Tehnologii folosite 42](#_Toc42678381)

[3.2.1 Android Studio 42](#_Toc42678382)

[3.2.2 Dependinte(Gradle File) 43](#_Toc42678383)

[3.3 Structura(Diagrama bloc explicata) 43](#_Toc42678384)

[3.3.1 Activitatea introductiva 43](#_Toc42678385)

[3.3.2 Activitatea de Log In 43](#_Toc42678386)

[3.3.3 Activitatea de Register 43](#_Toc42678387)

[3.3.4 Activitatea de Alegere a pozei 43](#_Toc42678388)

[3.3.5 Activitatea de obtinere a descrierei textuale 43](#_Toc42678389)

[Capitol 4 Firebase 44](#_Toc42678390)

[4.1 Motivatie/Descriere 44](#_Toc42678391)

[4.1.1 Ce este Firebase 44](#_Toc42678392)

[4.2 Autentificare 44](#_Toc42678393)

[4.3 Docker? 44](#_Toc42678394)

[4.4 Hosting?/Flask 44](#_Toc42678395)

[Capitol 5 Testare 45](#_Toc42678396)

[5.1 Descrierea Utilizarii 45](#_Toc42678397)

[5.2 Rezultate 45](#_Toc42678398)

[5.3 Scalabilitate 45](#_Toc42678399)

[5.4 Imbunatatiri 45](#_Toc42678400)

[Capitol 6 Concluzie 46](#_Toc42678401)

[Bibliografie 47](#_Toc42678402)

# Listă de figuri

[Figura 1‑1 Topologie aplicație 18](#_Toc42619096)

[Figura 2‑1 Machine Translation 23](#_Toc42619097)

[Figura 2‑2 Topologie model Image Captioning [5] 24](#_Toc42619098)

[Figura 2‑3 Convoluție [6] 25](#_Toc42619099)

[Figura 2‑4 ReLU 26](#_Toc42619100)

[Figura 2‑5 Imaginea prin straturile convoluționale [8] 26](#_Toc42619101)

[Figura 2‑6 Subeșantionare [6] 27](#_Toc42619102)

[Figura 2‑7 Max Pool [9] 27](#_Toc42619103)

[Figura 2‑8 Liniarizarea straturilor convoluționale [6] 28](#_Toc42619104)

[Figura 2‑9 Fully connected [6] 28](#_Toc42619105)

[Figura 2‑10 Rețea convoluțională [10] 28](#_Toc42619106)

[Figura 2‑11 LSTM 29](#_Toc42619107)

[Figura 2‑12 Rețea recurentă desfașurată in timp 30](#_Toc42619108)

[Figura 2‑13Topologii de rețele recurente [11] 30](#_Toc42619109)

[Figura 2‑14 Gated Recurrent Unit [13] 31](#_Toc42619110)

[Figura 2‑15 VGG16 [15] 34](#_Toc42619111)

[Figura 2‑16 Rezultate: Păsări 38](#_Toc42619112)

[Figura 2‑17 Rezultate: Atlet 39](#_Toc42619113)

[Figura 2‑18 Rezultate: Plajă 39](#_Toc42619114)

[Figura 2‑19 Rezultate: Urs 40](#_Toc42619115)

[Figura 2‑20 Rezultate: Barbat 40](#_Toc42619116)

[Figura 2‑21 Rezultate: Femeie și obiect 41](#_Toc42619117)

[Figura 2‑22 Rezultate: Femeie fără obiect 41](#_Toc42619118)

# Listă de Tabele

[Tabela 2‑1 Parametrii VGG16 35](#_Toc42632380)

[Tabela 2‑2 Model rețea recurentă 37](#_Toc42632381)

# Listă de acronime

**AI**: Artificial Intelligence(Inteligența artificială)

**ML**: Machine Learning

**UI**: User interface(Interfața grafică)

**Cuda**: Compute Unified Device Achitecture

**BaaS**: Backend-as-a-Service

**API**: Application Programming Interface (colecția de clase/funcții/interfețe cu ajutorul cărora exploatăm funcționalitatea unui framework sau a unei biblioteci)

**RNN**: Recurent Neural Network(Rețea neuronală recurentă)

**GRU**: Gated Recurrent Units

**LSTM:** Long-short term memory

**REST:** Representational state transfer

**ReLU:** Rectified linear unit

**IDE:** Integrated development environment

**AVD:** Android Virtual Device

# 

# Introducere

În zilele noastre dezvoltarea tehnologiei este mai rapidă iar acest lucru se poate observa mai ales în domeniul științei calculatoarelor. Inteligența artificială(AI) este la momentul de fața unul din vârfurile de lance în acest domeniu oferind o perspectivă diferită asupra rezolvării unor probleme care în programarea clasică s-au dovedit de-a lungul timpului dificil de abordat. Deși nu este un domeniu nou, el având bazele la începutul anilor 50 când metodele statistice care stau la baza algoritmilor de azi sunt găsite și perfecționate, implementarea lor și utilizarea acestor algoritmi in mediile de producție s-a dovedit dificilă până la începutul anilor 2000.

Apariția internetului și accesul la acesta pentru publicul larg la sfârșitul secolului 20 a fost unul din principalii factori ce au dus la reducerea în prim plan al inteligenței artificiale. Internetul rezolvă problema insuficienței datelor cu care cercetătorii în domeniu s-au lovit până în acel punct, facilitând crearea unor baze de date care ulterior să poată servi antrenării diferitelor modele matematice disponibile în acel moment.

Un alt lucru ce a împiedicat multă vreme ca acești algoritmi sa fie fezabili a fost lipsa puterii de procesare. A fost nevoie de zeci de ani pentru ca industria semiconductoarelor să ajungă la nivelul în care aceste modele să poate fi utilizate pe probleme complexe, cu un număr considerabil de clase, în scenarii utile aplicațiilor noastre de zi cu zi. În istoria recentă, Nvidia face parte din ceea ce se poate numi “Big bang-ul” rețelelor neuronale adânci, când în 2009 un astfel de model este antrenat pe o placă video sporind viteza de procesare a acestora de aproximativ 100 de ori.

În 2020 putem găsi componente AI în majoritatea programelor/aplicațiilor pe care noi le folosim de zi cu zi. Proiectul curent își propune realizarea un aplicații pe terminalul mobil care să folosească toate avansurile tehnologice menționate mai sus. Folosind un terminal mobil cu Android aplicația va fi capabilă sa facă poze sau să folosească poze deja existente în galeria telefonului cărora ulterior sa le adauge o descriere generată de un model de ML(machine learning). Aplicația va fi una te tip server-client unde, terminalul mobil(smatphone-ul) se va comporta ca un client pentru un server ce așteaptă să primească o poză de la telefon și ii va răspunde cu ieșirea modelului, o propoziție ce descrie acea poză.

Obiectivele pe care și le propune acest proiect sunt următoarele:

• Dezvoltarea unei aplicații Android

• Implementarea topologiei server client intre terminalele mobile și un server cu o putere de calcul îndeajuns de mare încât să ruleze modelul.

• Antrenarea modelului de rețea neuronală cară sa facă descrierea imaginilor

• Testarea aplicației

# Noțiuni teoretice necesare dezvoltării practice

## Topologia aplicației

Aplicația noastră are 3 componente software:

* O aplicație dezvoltată în Android Studio ce va servi ca front-end pentru aplicația noastră și va facilita utilizatorului un meniu intuitiv și ușor de folosit pentru a profita de facilitățile oferite.
* O componenta Firebase ce va servi ca back-end și pe care noi ne vom stoca modelul antrenat pentru a genera descrierea imaginilor.
* Un model de rețea neuronală adâncă dezvoltat in Python cu ajutorul librăriei Tensorflow.

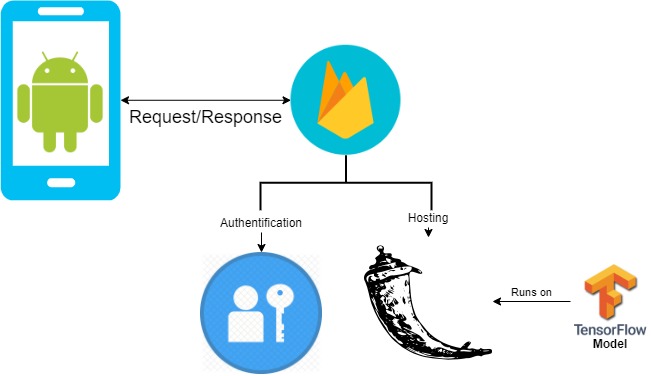


Figura ‑ Topologie aplicație

Modelul obținut de noi în urma antrenării va fi prea mare pentru a putea fi stocat pe terminalul mobil. De asemenea nu orice terminal mobil are posibilitatea de a rula modele complexe pe arhitectura lor hardware așa că pentru a acomoda aplicația pe cat mai multe terminale mobile s-a luat decizia asupra acestei arhitecturi.

## Terminalul mobil

Aplicația poate fi instalată și va rula pe orice telefon mobil, tabletă sau dispoziiv ce rulează o distribuție de android mai noua decăt Nougat 7.0 și care dispune de o camera foto și de conexiune la internet. Cerințele acestea vor fi minimale întrucât marea parte a puterii de procesare va fi oferită de server. Pentru a testa aplicația noiastra am utilizate 2 terminale mobile: un Samsung Galaxy S9 și un Google Pixel A3(folosit într-un emulator).

### Platforma Android

Android este o platformă software și un sistem de operare pentru dispozitive și telefoane mobile bazată pe nucleul Linux, dezvoltată inițial de compania Google, iar mai târziu de consorțiul comercial Open Handset Alliance. Android permite dezvoltatorilor să scrie cod gestionat în limbajul Java și începând cu anul 2017 Kotlin controlând dispozitivul prin intermediul bibliotecilor dezvoltate de Google. Aplicațiile scrise în C și în alte limbaje pot fi compilate în cod mașină ARM și executate, dar acest model de dezvoltare nu este sprijinit oficial de către Google. [1]

Deși Java a fost limbajul de programare care a dominat o buna perioadă de timp dezvoltarea aplicațiilor pe platforma Android, dezvoltatorii au posibilitatea să folosească Kotlin care în mai puțin de 3 ani de la integrarea în platforma Android a ajuns sa fie limbajul oficial preferat de Google pentru Android. Kotlin face ca dezvoltarea aplicațiilor sa fie mai concisă (mai mult cu mai puține linii de cod) și elimină necesitatea dezvoltatorilor să lege fiecare element din UI(user interface) cu codul din spate. Fiind un limbaj nou, el oferă un nivel de abstractizare similar cu cel al JavaScript-ului sau Python-ului. Cu toate acestea el se poate integra ușor cu Java, dezvoltatorul având posibilitatea să folosească in Kotlin instanțe ale claselor scrise în Java și viceversa. [2]

### Hardware

În momentul de față pe majoritatea terminalelor mobile pe lângă să asigure o comunicația audio fără fir sunt capabile să facă o multitudine de alte sarcini prin intermediul componentelor adiționale pe care le regăsim în formatul lor compact. Display-urle bazate pe tehnologii LED sunt capabile să redea imagini de calitate la rezoluții înalte și cu o gama largă de culori, procesoarele, memoria și stocarea sunt capabile să susțină sisteme de operare comparabile ca performanțe cu cele regăsite pe calculatoare, modem-urile integrate asigură comunicații de viteză și viteze înalte de transfer pe internet. Camerele foto integrate dispun de caracteristici precum: autofocus, senzor de lumină, flash, stabilizare de imagine digitală sau mecanică. De asemenea putem regăsi senzori adiționali precum: accelerometru, giroscop, busolă digitală, senzori de proximitate, magnetometru etc.

Terminalele mobile de ultimă generație dispun de acceleratoare grafice cu arhitecturi CUDA, arhitectura utilizată și pe plăcile video regăsite pe calculatoare. Aceste acceleratoare permite rularea unor redarea unor jocuri/aplicații cu înaltă calitate grafică și pun la dispoziție o suită de biblioteci de software de ML accelerată complet pe GPU.

## Firebase

Firebase este un backend ca un serviciu (BaaS). Firebase elimină necesitatea dezvoltatorului de a se concentra pe dezvoltarea și managementul unui server, și pune la dispoziție dezvoltatorului un API generic pentru a ii ușura munca. Printre facilitățile oferite de Firebase noi vom utiliza următoarele:

* Firebase Analytics pentru a obține statistici despre folosirea aplicației
* Firebase Auth pentru a permite utilizatorilor sa se autentifice pe un acont propriu și pentru a le proteja pozele trimise către server
* Firebase Hosting pentru a ține scripturile în Python
* ML Kit pentru a integra modelul antrenat

In momentul de față, Firebase este o soluție populara atât pentru aplicațiile Android si IOS cât și pentru

aplicațiile web oferind atât o interfață stabilă cu backend-ul aplicațiilor cat și metrici relevante legate de utilizarea aplicatiei.

## Python

Python este un limbaj de programare interpretat, la nivel înalt, cu scop general. Creată de Guido van Rossum și lansată pentru prima dată în 1991, filozofia de proiectare a lui Python subliniază lizibilitatea codurilor prin utilizarea spațiilor pentru a delimita diferite structuri. Abordarea orientată spre obiect și multitudinea de funcții si librarii cu care el vine deja instalat face ca Python să fie in ziua de azi unul din limbajele de programare preferat pentru scriptare, proiecte de dimensiuni mici sau back-end în aplicații. Suportă mai multe paradigme de programare, incluzând programarea procedurală, orientată pe obiecte și funcționale

Pe lângă librăriile cu care python vine deja instalat, în cadrul proiectului nostru vom mai utiliza, pe lângă Tensorflow, și alte librarii/pachete precum:

* NumPy: o librărie care oferă suport pentru operațiile cu tensori, și un set de funcții pe care le putem folosi pe aceștia
* Matplotlib: librărie ce permite plotarea datelor in grafice
* OpenCV: o bibliotecă de funcții informatice specializată pe vedere pe care o vom folosi în prelucrarea pozelor
* Pandas: o librărie folosită pentru manipularea datelor
* Flask: un micro-framework pentru dezvoltare WEB

## Flask

Flask a fost creat de Armin Ronacher de la Pocoo, un grup internațional de pasionați Python format în 2004. Deși inițial proiectul a fost dezvoltat ca o glumă de 1 Aprilie, acesta avea sa devina o alternativa solida pentru Django in dezvoltarea aplicațiilor web. Simplitatea și numărul de linii de cod redus au dus la popularitatea framework-ului de azi în aplicații bazate pe microservicii. Titulatura de microframework se datorează faptului ca nu necesita alte librării sau software-uri pentru a funcționa.

Caracteristicile framework-ului:

* Permite crearea de servere de debug cât și de producție
* Permite testarea modulelor separate(Unit Testing)
* Oferă suport pentru API-uri de REST
* Asistență pentru cookie-uri sigure, pentru sesiunile din partea clientului
* Extensii disponibile pentru a îmbunătăți caracteristicile dorite, cum ar fi suport pentru manipularea bazelor de date [3]

## Tensorflow

TensorFlow este o bibliotecă software gratuită și open-source pentru flux mare de date și programare tensorială pentru o gamă largă de sarcini. Este o bibliotecă simbolică de matematică și este, de asemenea, utilizată pentru aplicații de învățare automată, cum ar fi rețelele neuronale. Este utilizat atât pentru cercetare, cât și pentru producție la Google.‍ Librăria este disponibilă pentru mai multe tehnologii de programare cum ar fi C, Java, Python și Javascript însă este mai stabilă pe ultimele 2. In cadrul proiectului nostru vom utiliza varianta librăriei din Python alături Keras.

Partea funcțională a TF este implementată în C++ și CUDA, în timp ce API-ul cel mai frecvent utilizat este creat în limbajul Python. Așadar, în TF se programează folosind Python, dar procesarea efectivă se realizează de către un engine dezvoltat in C++ / CUDA. [4]

### Modul de funcționare

Un program TF are două componente esențiale:

* un model, ce conține totalitatea operațiilor ce se doresc a se efectua, precum și a datelor (tensorilor) ce se doresc a fi determinate. Modelul se implementează folosind o structură de tip graf, ce conține succesiunea de execuție a operațiilor și de evaluare a tensorilor (i.e. de determinare a valorilor variabilelor modelului).Acest graf este realizat conform principiului de programare data flow , de unde și denumirea de TensorFlow – o bibliotecă ce prelucrează tensori folosind operații structurate pe principiul data flow (pentru simplificare, prin tensor înțelegem un array multidimensional – scalar, vector, matrice 2D, 3D etc.) [4]
* sesiune – obiect ce permite execuția parțială sau completă a modelului menționat anterior. În cadrul unei sesiuni se rezolvă problema dorită cu ajutorul modelului definit în prealabil. [4]

### Keras

Este API-ul la nivel înalt al TensorFlow-ului pentru construirea și antrenarea modelelor de ML. Este folosit pentru prototipare rapidă, cercetare de ultimă generație și producție, cu trei avantaje cheie:

* Ușor de utilizat: Keras are o interfață simplă, consistentă, optimizată pentru cazuri de utilizare obișnuită. Oferă feedback clar și acționabil pentru erorile utilizatorilor.
* Modulară și compozibilă: Modelele Keras sunt realizate prin conectarea blocurilor de construcții configurabile, cu câteva restricții.
* Ușor de extins: Scrierea blocurilor de construcții personalizate pentru a exprima noi idei de cercetare. Creați noi straturi, valori, funcții de pierdere și dezvoltați modele de ultimă generație.

# Modelul Rețelei Neuronale

Descrierea conținutului vizual (Image captioning) își propune descrierea conținutului vizual generând un text. La baza acestei idei a stat ideea de traducere bazată pe ML. Acolo întâlneam o arhitectura de tip autoecoder, unde prima jumătate a rețelei (encoder-ul) coda textul prezent la intrare într-un format cunoscut ca și vector-gând iar cea de-a doua parte a rețelei decoda semantica acestui vector într-o limbă noua. Acest vector-gând reprezintă semantica propoziției și permite ca aceasta sa fie decodat în altă limbă într-o propoziție a cărei lungime nu trebuie sa fie egala cu cea de la intrare, dar care păstrează informația primită. Atât encoder-ul cât si decoder-ul erau rețele neuronale recurente.

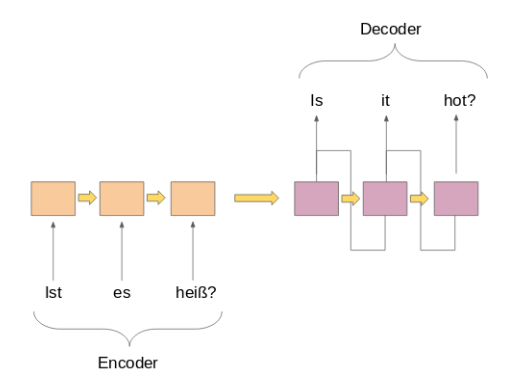


Figura ‑ Machine Translation

Plecând de la aceeași idee vom înlocui encoder-ul cu un model de rețea neuronala convoluțională capabilă sa recunoască obiectele din imagine și care la ieșire va scoate un vector-gând similar cu cel prezent în arhitectura de traducere. Pentru a obține acest vector vom elimina din rețea ultimul strat de clasificare și vom redirecționa ieșirea penultimului strat către intrarea rețelei neuronale recurente. Cu toate acestea, dimensiunea internă a RNN-ului este de numai 512, de aceea avem nevoie de un strat intermediar complet (conectat) dens pentru a face maparea vectorului cu 4096 de elemente la un vector cu doar 512 elemente.

Decodorul folosește apoi acest vector împreună cu un marker de pornire „ssss” pentru a începe să producă cuvinte de ieșire. În prima iterație, este posibil să dea naștere cuvântului „big”. Apoi introducem acest cuvânt în decoder și sperăm că vom scoate cuvântul „brown” și așa mai departe. În cele din urmă, am generat textul „big brown bear sitting eeee" unde „eeee” marchează sfârșitul textului.

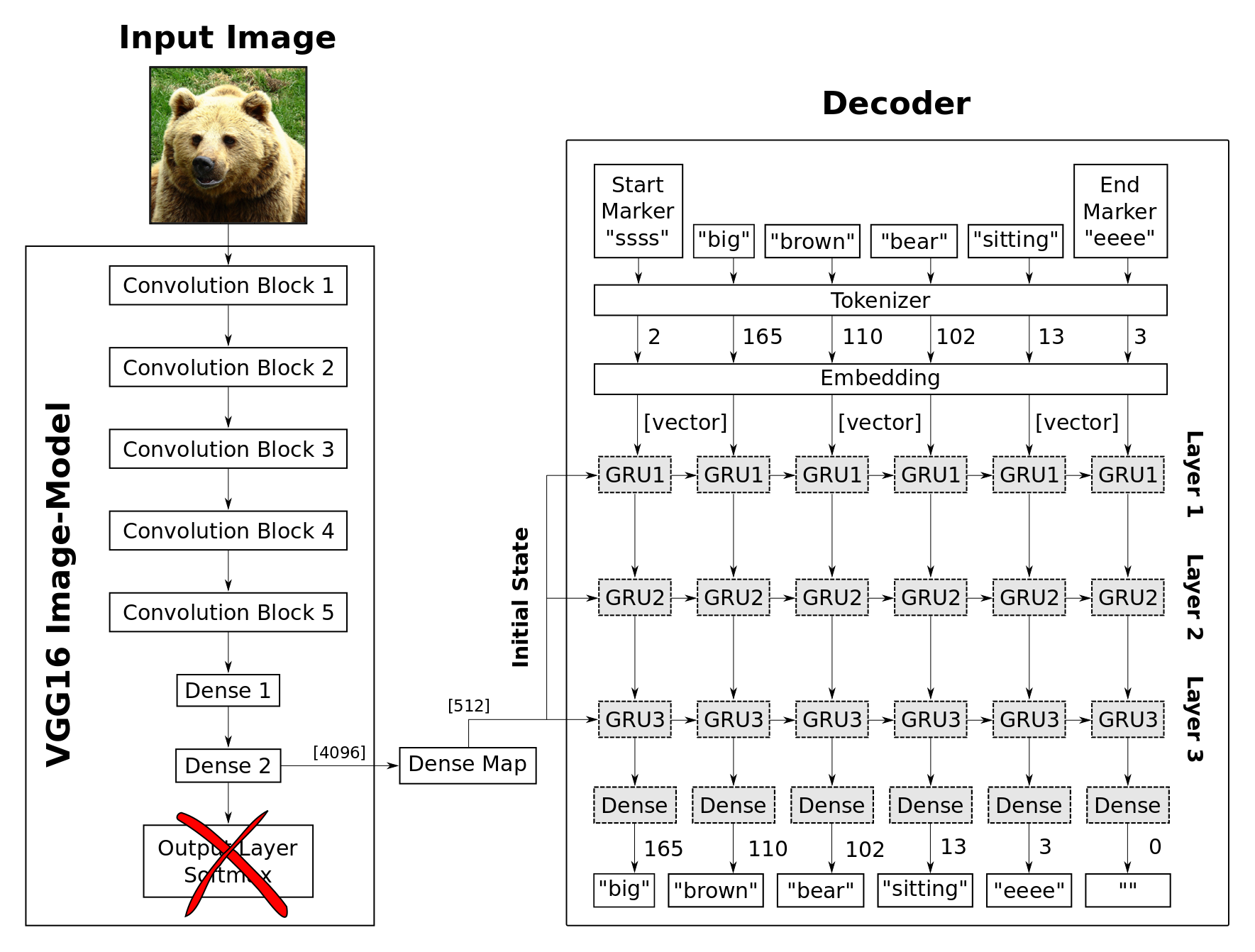


Figura ‑ Topologie model Image Captioning [5]

Întrucât cele 2 arhitecturi diferă, VGG16 fiind unidirecțională(feed-forward), iar decoderul fiind o arhitectură recurentă, le vom antrena separat. Procedeul de antrenare este unul supervizat, în cazul rețelei convoluționale folosind “COCO data-set’’ ce conține multe poze cu descrierile lor aferente, iar output-ul primei rețele va fi codat(tokenized). În cazul rețelei recurente alcătuită din 3 straturi GRU(Gated Recurrent Units), vom aplica același procedeu folosind descrierea pozelor codată in același format în care el o va primi in folosință de la rețeaua convoluțională.

## Rețelele convoluționale

Este o arhitectură de rețea neuronală a capabilă sa extragă trăsături din imaginile propagate de-a lungul ei. Primele straturi convoluționale vor extrage trăsături de finețe, urmând ca pe măsură ce rețeaua se adâncește la nivelul unui strat sa putem observa trăsături ce descriu forme geometrice sau chiar obiecte de mici dimensiune. Acest tip de rețele au 3 straturi distincte:

**Stratul convoluțional**: este stratul de baza al rețelelor neuronale, fiind constituit dintr-o serie de filtre ce acoperă o mică porțiune din imagine. Operația matematica specifică este suma produsului element cu element intre fereastra acoperita de filtru și filtrul însăși. Filtrul cu care este aplicată convoluția pe o matrice este deplasat pe suprafața imaginii din colțul din stânga sus la dreapta și mai apoi pe liniile următoare până când ajunge in colțul din dreapta jos, obținându-se o noua matrice de valori.

Stratul convoluțional are proprietatea de a găsi caracteristici în imagine indiferent de poziția lor în imagine, datorită conectivității locale și a ponderilor comune pe fiecare porțiune a imaginii.

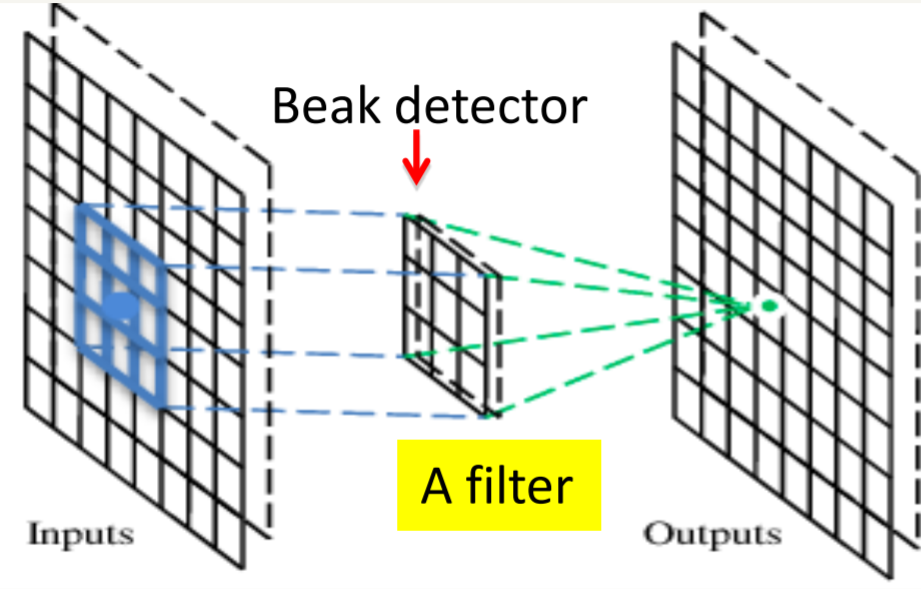


Figura ‑ Convoluție [6]

O rețea convoluțională reprezintă o succesiune de straturi convoluționale intercalate cu funcțiile de activare. In mod uzual este utilizată funcția de activare ReLU(rectified linear unit). Comportamentul acestei funcții este descris astfel:

 .

Funcția are un comportament similar cu modul de activare al neuronilor noștri și s-a dovedit a avea rezultate mai bune in faza de antrenare a modelelor de rețele neuronale datorită propagării mai eficiente a gradientului. Un alt avantaj ar fi simplitatea computaționala a acestei funcții comparată cu alte funcții precum *tanh* sau *sigmoida.*

Dezavantajele utilizării sale sunt faptul ca nu este o funcție centrată in 0, că deși este o funcție invariantă cu un coeficient, în procesul de antrenare(backpropagation) funcția nu este derivabila pe intervalul  și nu in ultimul rând, neuronii pot fi împinși într-o stare inactiva in procesul de antrenare. Alte variații: Noisy ReLU, Leaky ReLU, Parametric ReLU, ELU, încearcă să rezolve câteva din problemele menționate anterior. [7]

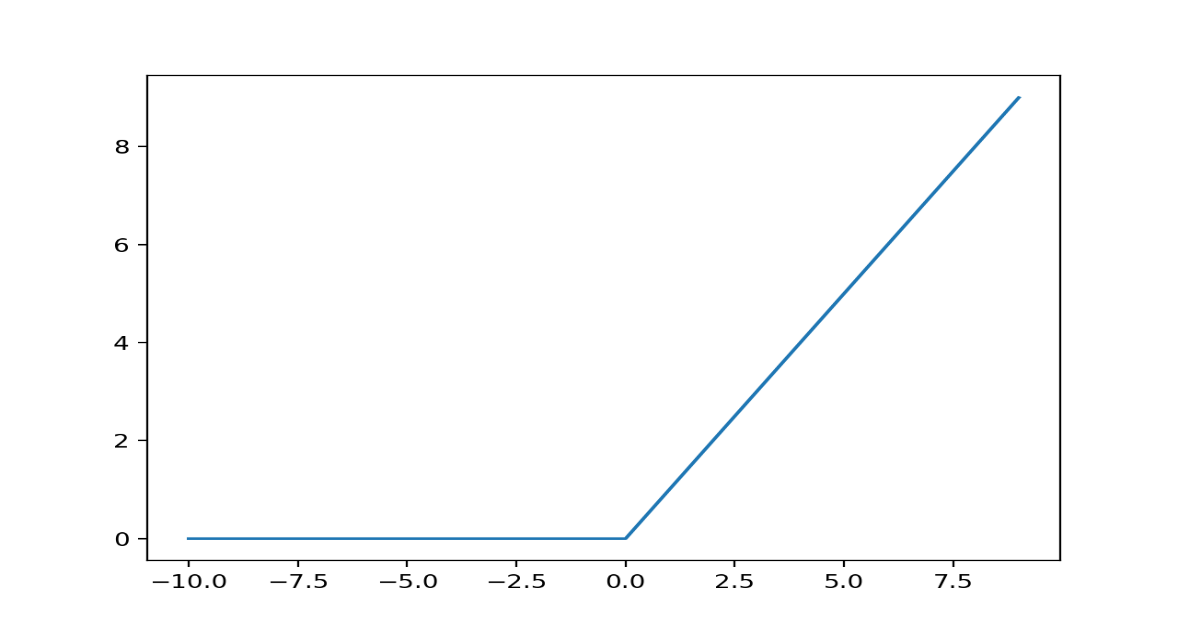


Figura ‑ ReLU

Într-o înșiruire de straturi convoluționale, primele straturi vor fi capabile sa determine trăsături fine, contururi mici și muchii, iar pe măsură ce modelul devine din ce în ce mai adânc, putem observa trăsături de nivel înalt precum și chiar obiecte de dimensiune mai mică.

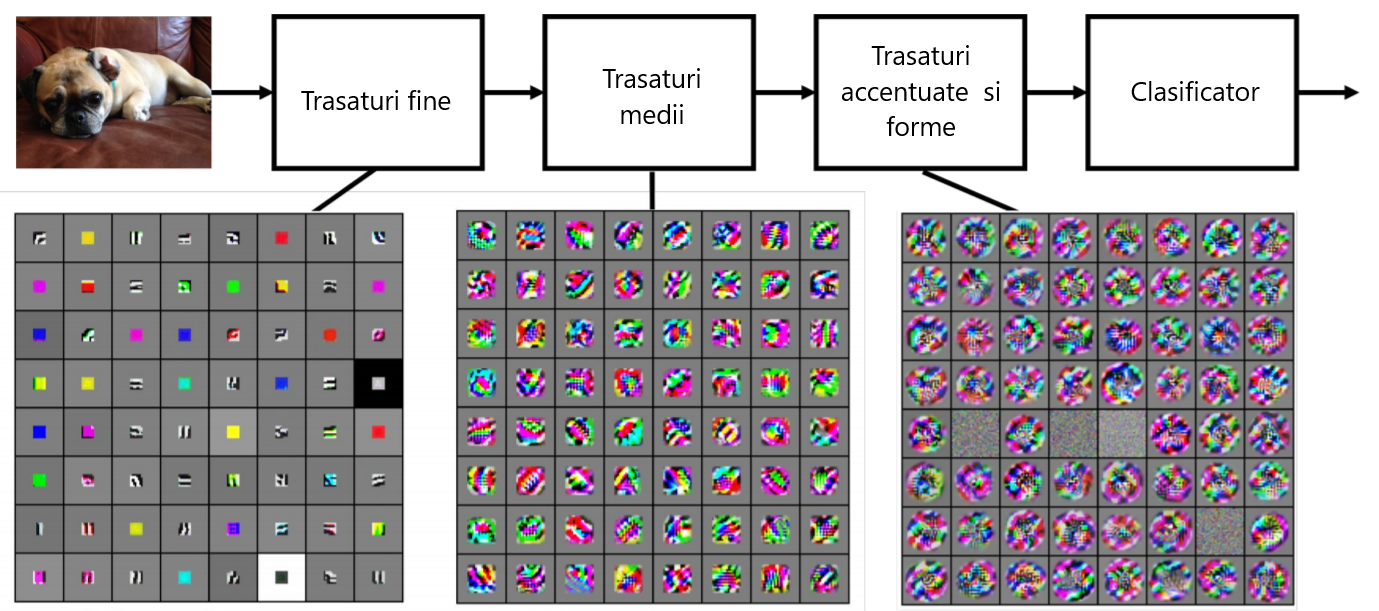


Figura ‑ Imaginea prin straturile convoluționale [8]

**Straturi de pooling** al căror scop este să subeșantioneze informația in diferite stagii ale rețelei. Această subeșantionare nu va distorsiona informația din imagine, dar va reduce numărul de pixeli/calitatea imaginii pentru a putea fi mai ușor de lucrat cu aceasta.



Figura ‑ Subeșantionare [6]

De regulă aceste straturi sunt prezente într-o rețea la fiecare 1-2 straturi convoluționale și reduc la jumătate informația aflată în imagine la acel moment fie prin:

* Max-pooling-maximul dintr-o fereastră(2x2 de regulă)
* Avg-pooling- media dintr-o fereastră

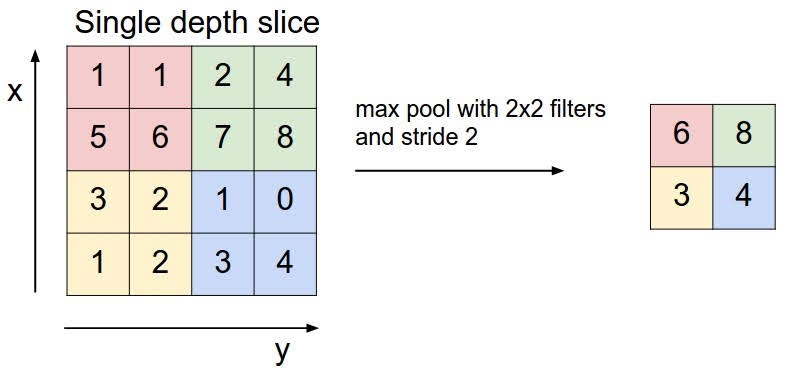
****

Figura ‑ Max Pool [9]

**Straturi Dense/Fully connected** pe ultimele straturi dintr-o rețea convoluțională vom găsi straturi dense (fully connected) care preiau trăsăturile, determinate de straturile convoluționale și clasifică imaginile. Din punct de vedere computațional straturile dense trebuie să rețină un volum de ponderi semnificativ mai mare decât un strat convoluțional unde ponderile sunt refolosite pe fiecare fereastră. Intre straturile convoluționale și acestea mai exista un strat de liniarizare a informație care liniarizează imaginile venite din straturile convoluționale pentru a putea fi procesate de straturile fully connected. Ultimul va un număr de neuroni egal cu numărul de clase după care se dorește clasificarea.

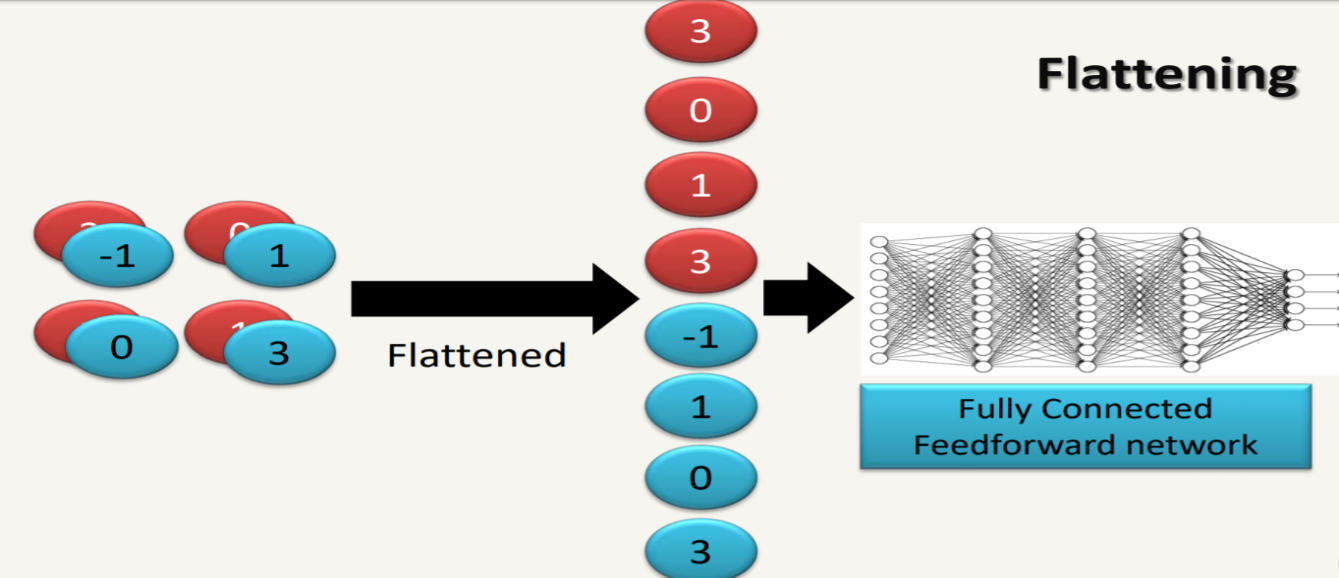


Figura ‑ Liniarizarea straturilor convoluționale [6]

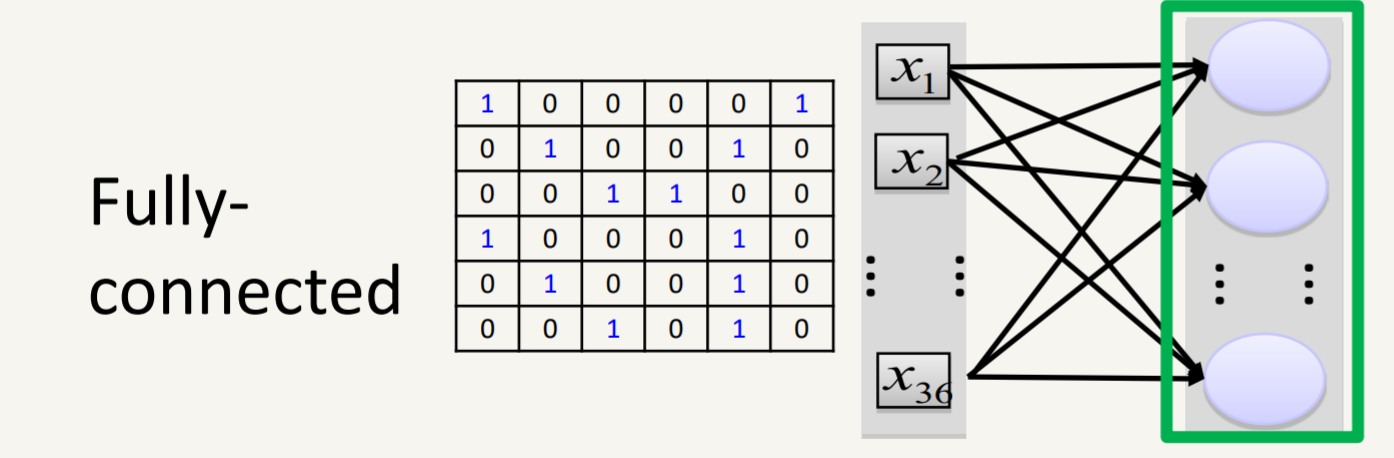
****

Figura ‑ Fully connected [6]

Cele 3 tipuri de straturi menționate anterior se combină in ordinea specificată pentru a forma diferite arhitecturi de rețele neuronale convoluționale după cum se poate observa și in imaginea de mai jos.

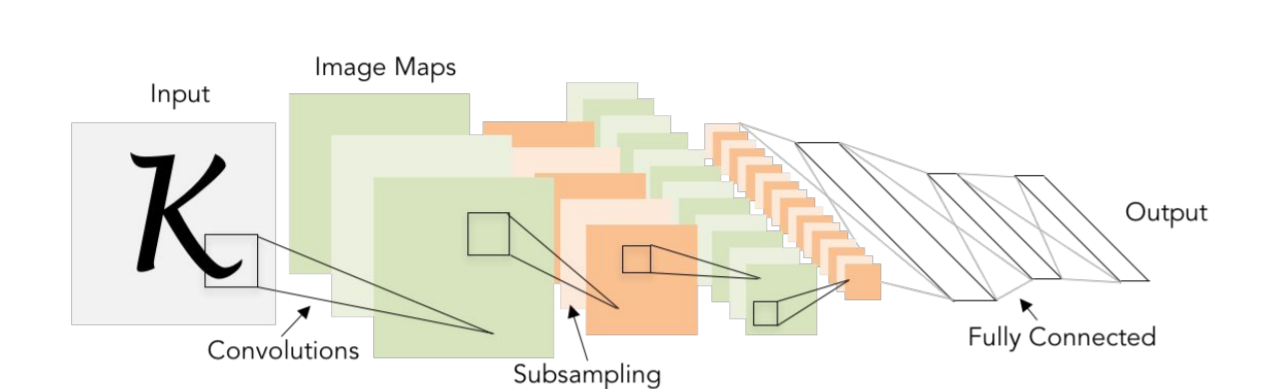


Figura ‑ Rețea convoluțională [10]

## Rețele recurente

Dacă o rețea feed forward, dens conectată primea la intrare ei toată informația, o rețea recurentă primește bucăți din informație la momente diferite de timp. Rețeaua e capabilă sa proceseze informația(cuvintele in cazul nostru) pe rând iar la momentul următor de timp, când aceasta va primi următorul cuvânt, să il proceseze, folosind aceleași ponderi și ținând cont de cuvintele pe care le-a primit anterior pentru a produce o ieșire care să clasifice întreaga secvență primită. Privită pe toată durata s-a de timp, adică pe lungimea vectorului de intrare(număr de cuvinte) ea poate fi convertită ca o rețea neuronală densă care după fiecare strat primește încă o bucată din informația de intrare, lucru ce se poate vedea și in Figura 2‑12 Rețea recurentă desfășurată in timp.

Asemenea arhitecturi s-au dovedit utile în lucrul cu secvențele audio sau video dar și in modele care au ca scop traducerea dintr-o limba în alta. Sunt capabile să recunoască tipare în datele cu care lucrează, fapt ce s-a dovedit util atunci când încercăm să traducem dintr-o limbă în alta, aceeași idee fiind exprimată cu un număr diferit de cuvinte de la limbă la limbă. Similar proceselor de traducere și recunoaștere vocală, încercăm sa obținem o ‘traducere’ a pozelor de la intrarea modelului nostru interpretându-le inițial cu un model de rețea convoluțională ce va obține o secvența similara cu a celora obținute de encoder-ul arhitecturilor de recunoaștere vocală sau traducere dintr-o limbă in alta.

Principala problemă a acestor rețele este dată de lungimea secvenței. O dată cu trecerea timpului primele cuvinte care au intrat în rețea vor avea o pondere mai mică asupra rezultatului final fața de ultimele cuvinte. Această problemă se va reflecta și in procedeul de antrenare unde gradienții vor fi din ce in ce mai mici pe măsură ce ne întoarcem în timp pentru a modifica ponderile neuronilor.

Soluții precum LSTM(long-short term memory) și GRU au fost găsite pentru a păstra relevantă informația din pașii anteriori. Aceste unități sunt capabile să rețină informații de la aproximativ 1000 de cuvinte.

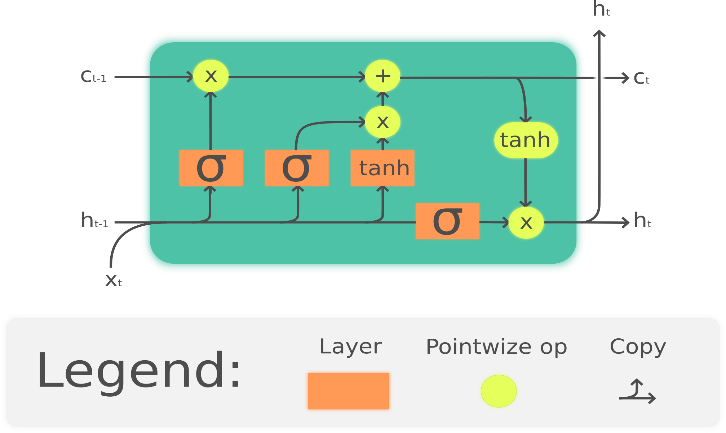


Figura ‑ LSTM

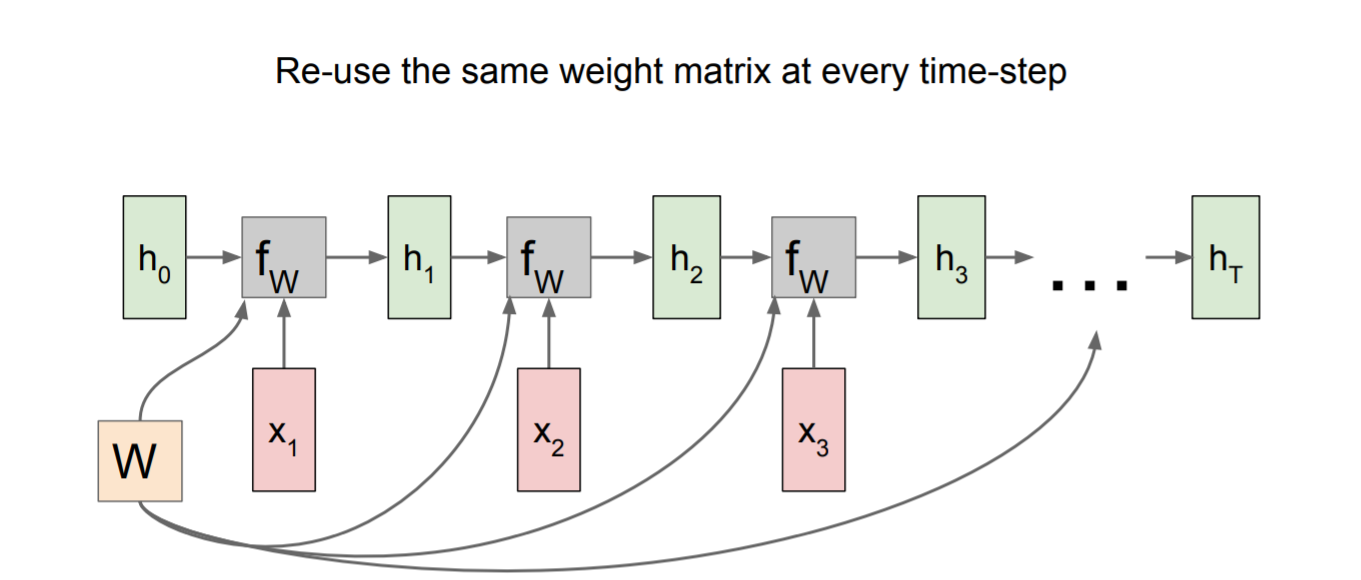


Figura ‑ Rețea recurentă desfășurată in timp

În funcție de topologia rețelelor acestea pot avea diverse aplicații după cum se poate observa și in figura de mai jos.

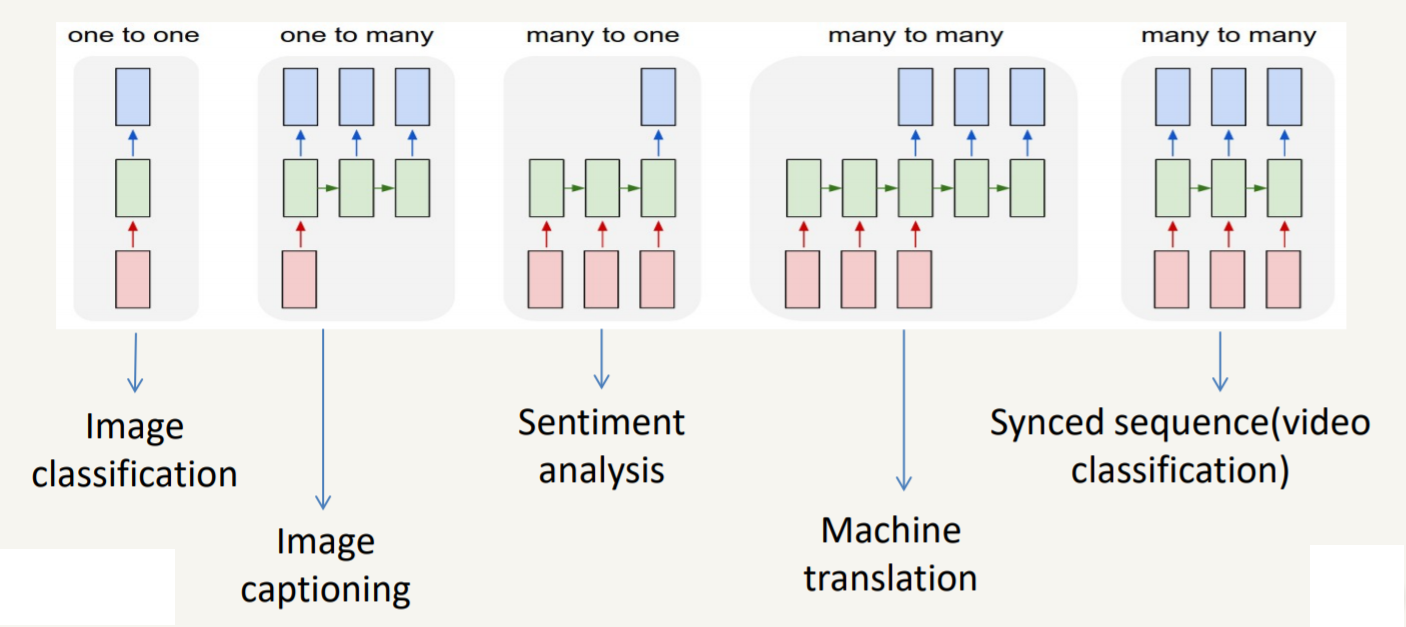


Figura ‑Topologii de rețele recurente [11]

### Gated recurrent unit

Gated recurrent units (GRUs) este mecanismul de neuron recurent pe care noi îl vom utiliza in descrierea conținutului vizual, în stagiul de decodare. Descoperiți in 2014 de  Kyunghyun Cho [12] structura lor este similară cu cea a LSTM-urilor, având o poartă de uitare, pentru a putea reține doar informațiile relevante, însă are mai puțini parametrii ca acesta și nu are o poartă de ieșire. Deși nu la fel de puternic în aplicațiile de traducere precum LSTM, GRU au obținut performanțe similare cu acestea în recunoaștere audio și de discurs.

**Modelul matematic**

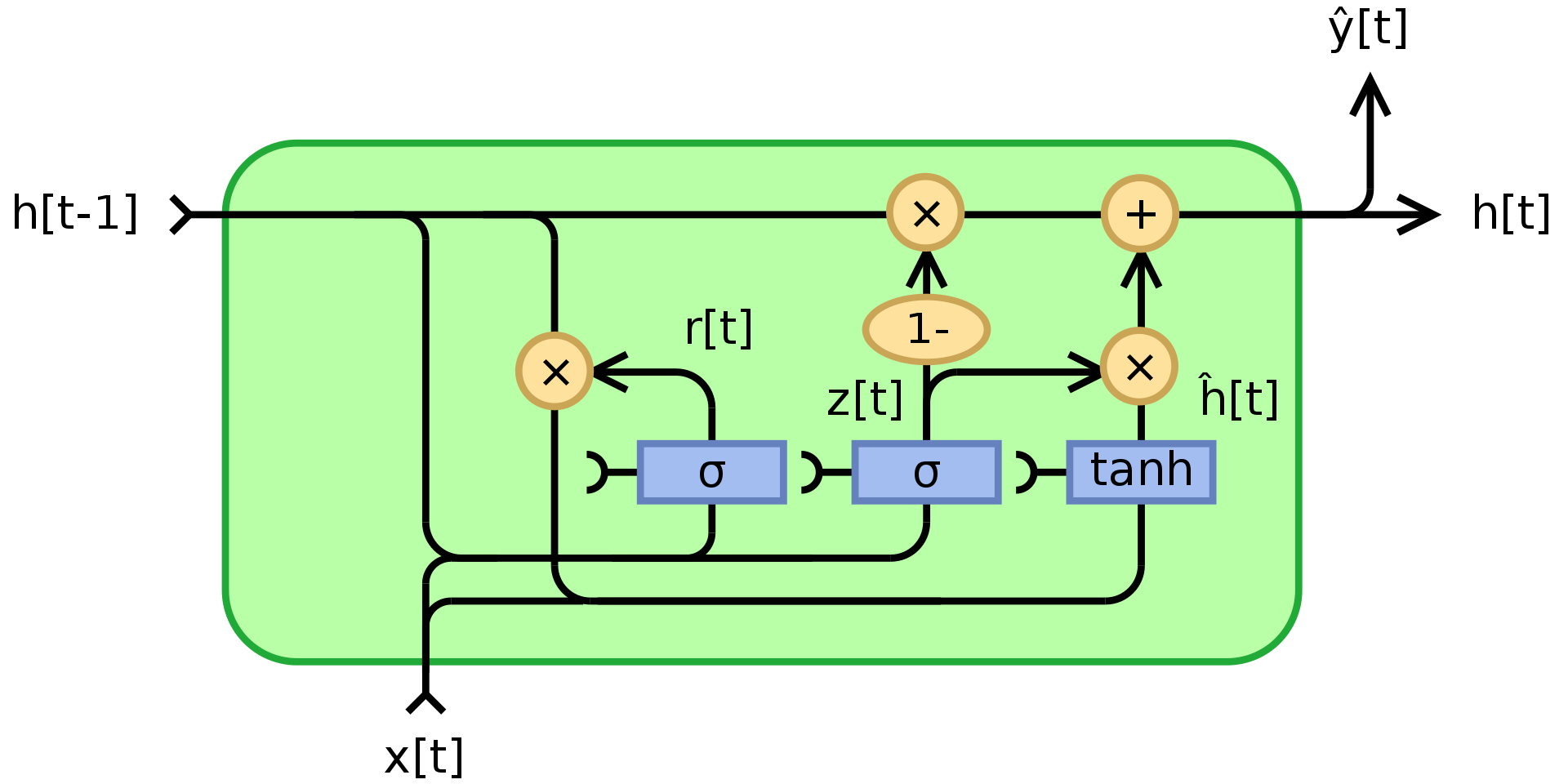
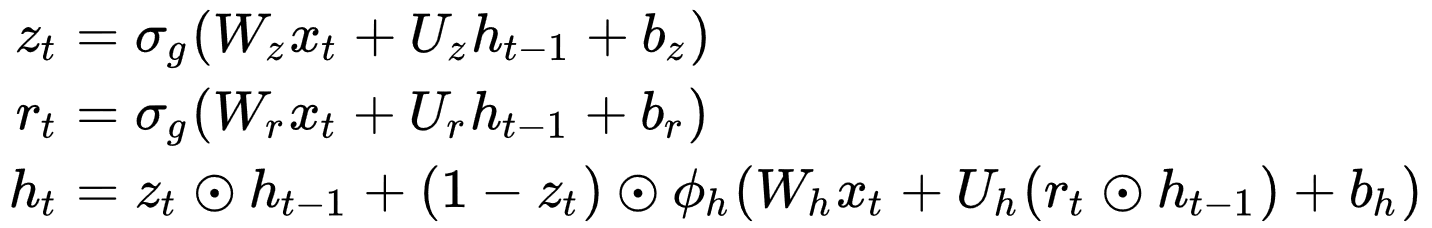
****

Figura ‑ Gated Recurrent Unit [13]

La t=0, ieșirea 



Unde,

*  -vectorul de intrare
*  -vectorul de ieșire
*  -vectorul actualizat
*  - vectorul de resetare al informației
*  și  parametrii, matricea și vectorul

Funcțiile de activare

*  - Funcția de activare sigmoida
*  - Funcția de activare tangentă hiperbolică

### Straturile de embedding

Straturile de embedding(embedding layers) oferă o reprezentare densă a cuvintelor sub forma unui vector de valori întregi. Acest tip de strat a apărut ca o îmbunătățire a reprezentării baga de cuvinte(bag of words) datorită simplității sale. Valorile vectoriale atribuite fiecărui cuvânt poate fi învățate din texte și refolosite de la un proiect la altul. Pot fi și învățate în cadrul procesului de antrenare.

Spre deosebire de “bag of words”, straturile de embedding învață cuvintele în raport cu vecinii lor uzuali, iar, pe un vocabular de aceeași lungime, vectorii de reprezentare pentru un cuvănt au dimensiuni mai mici decât vectorii folosiți în algoritmul „bag of words”.

Este un strat flexibil care poate fi utilizat în mai multe moduri, cum ar fi:

* Poate fi folosit singur pentru a învăța o încorporare de cuvinte care poate fi salvată și folosită într-un alt model ulterior.
* Poate fi folosit ca parte a unui model de învățare adâncă, în care încorporarea(embedding-ul) este învățată împreună cu modelul însuși.
* Poate fi folosit pentru a încărca un model pre-instruit de încorporare a cuvintelor, un tip de învățare prin transfer.

În KERAS, când lucram cu un astfel de strat trebuie sa specificăm 3 parametri:

* input\_dim: Aceasta este dimensiunea vocabularului de cuvinte. De exemplu, dacă datele sunt întregi codificate la valori cuprinse între 0-10, atunci dimensiunea vocabularului ar fi de 11 cuvinte.
* output\_dim: Aceasta este dimensiunea spațiului vectorial în care vor fi încorporate cuvintele. Definim dimensiunea vectorilor de ieșire din acest strat pentru fiecare cuvânt.
* input\_length: Aceasta este lungimea secvențelor de intrare. De exemplu, dacă toate documentele noastre ar fi alcătuite din 1000 de cuvinte, acesta ar fi 1000.

## Transfer learning

Transfer learning (Învățarea prin transfer) se referă la folosirea unor cunoștințe deja acumulate prin experiențe anterioare sau probleme deja rezolvate pentru a găsi soluția unor probleme noi. În contextul ML ea face referirea la utilizarea unor modele deja antrenate pentru a rezolva probleme pe seturi de date noi. [14]

Oamenii pot adesea sa își folosească cunoștințele deja existente despre un domeniu pentru a rezolva probleme dintr-un domeniu adiacent cu acesta:

* Șah -> Dame
* Matematică -> Informatică
* Fotbal -> Rugby
* Patinaj -> Hochei

Cazuri în care am dori sa plecăm de la un model de rețea deja antrenată pentru a rezolva o problemă:

* Datele din setul nostru de antrenare nu sunt etichetate, așa ca folosim un model deja antrenat pentru a le grupa in clasa corespondentă.
* Detecția de obiecte. Poate ca modelul de la care am plecat nu era antrenat să detecteze clasele de care noi avem nevoie, însă putem reantrena ultimele straturi pentru a determina clasele de interes (Exemplu: Image Matting unde avem 2 clase fundalul și persoana). Formele pe care rețeaua noastră știe deja sa le detecteze, aflate in primele straturi ale rețelei noastre, pot fi reutilizate pentru a câștiga timp și a suplimenta pentru un volum de date de antrenare mai mic.
* Transferul de trăsături folosit în rezolvarea unei probleme.

## Baza de date

Pentru a antrena modelul nostru am utilizat baza de date COCO(Microsoft COCO: Common Objects in Context). COCO este un set de date de detecție, segmentare și descriere a obiectelor la scară largă. COCO are mai multe caracteristici: obiectel segmentate, segmentarea materialelor superpixel, 330 de mi de imagini din care mai mult de 200 de mii etichetate, 91 de categorii de obiecte și peste 1,5 imagini per tip de obiect, cel puțin 5 descrieri pe o imagine. Scopul acestei baze de date a fost sa ajute la rezolvarea unei probleme actuale in domeniul ML, și anume înțelegerea șcenelor.

Pentru construirea acestei baze de date, echipa formată din:Tsung-Yi, Lin Michael, Maire Serge, Belongie Lubomir, Bourdev Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollar au utilizat un instrument numit Amazon Mechanical Turk. Instrumentul le-a permis acestora adnotarea unui volum mare de date prin angajarea mai multor persoane.

Pentru a culege volumul mare de poze aceștia au efectuat căutări de baze de date interogând după tipuri de scene. În 328,000 de poze s-au etichetat aproximativ 2,5 milioane de obiecte. Comparativ cu Imagenet, COCO are mai puține clase însă are un număr mai mare de poze per clasă lucru ce poate fi util in problemele în care localizarea obiectelor în imagine este obiectivul principal. Numărul mare de imagini per clasă cat și numărul mare de descrieri textuale per imagine a reprezentat criteriul principal de selecție pentru sarcina de descriere conținutului vizual.

## Implementare și Antrenare

Am încercat 2 medii pentru a antrena modelul de descriere a imaginii. Primul este laptopul personal ce dispune de o memorie de 16 GB, un procesor i7-8750H și o placa video Nvidia Geforce GTX 1060 de 6GB. Cel de al doilea este o instanță de GPU in CodeColab. Problemele întâmpinate cu cea din urma opțiune au fost datorate dimensiunii mari a bazei de date (40 GB arhiva+ date) și a duratei reduse de viață a instanței. Pentru a putea încărca baza de date de fiecare dată aș fi avut nevoie de un cont premium de google drive pentru a nu descărca și prelucra mereu setul de date. Altă problema a fost că durata unei epoci, în funcție de parametrii de antrenare era cuprinsă între 2 și 6 ore. Am preferat să utilizez resursele laptopului meu pentru a face antrenarea modelului.

Primul pas în antrenarea rețelei noastre a fost reprezentat de descărcarea setului de date. Pentru antrenare am avut 118287 de imagini. Pixelii imaginilor au fost scalați între 0 și 1. Pentru partea de encodare a rețelei am utilizat un model de VGG16 antrenat pe ImageNet.

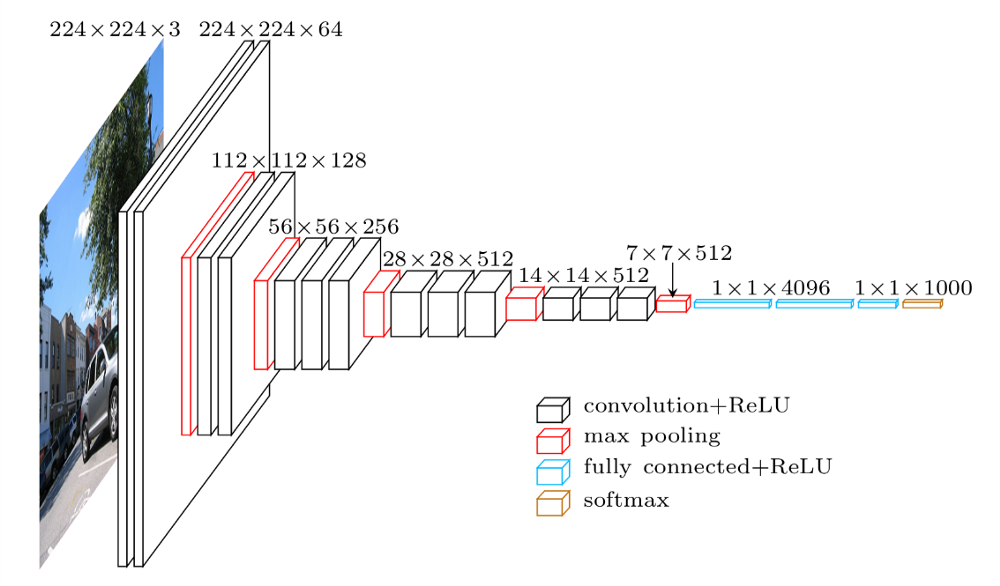


Figura ‑ VGG16 [15]

Model: "vgg16"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

input\_1 (InputLayer) [(None, 224, 224, 3)] 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block1\_conv1 (Conv2D) (None, 224, 224, 64) 1792

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block1\_conv2 (Conv2D) (None, 224, 224, 64) 36928

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block1\_pool (MaxPooling2D) (None, 112, 112, 64) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block2\_conv1 (Conv2D) (None, 112, 112, 128) 73856

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block2\_conv2 (Conv2D) (None, 112, 112, 128) 147584

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block2\_pool (MaxPooling2D) (None, 56, 56, 128) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block3\_conv1 (Conv2D) (None, 56, 56, 256) 295168

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block3\_conv2 (Conv2D) (None, 56, 56, 256) 590080

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block3\_conv3 (Conv2D) (None, 56, 56, 256) 590080

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block3\_pool (MaxPooling2D) (None, 28, 28, 256) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block4\_conv1 (Conv2D) (None, 28, 28, 512) 1180160

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block4\_conv2 (Conv2D) (None, 28, 28, 512) 2359808

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block4\_conv3 (Conv2D) (None, 28, 28, 512) 2359808

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block4\_pool (MaxPooling2D) (None, 14, 14, 512) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block5\_conv1 (Conv2D) (None, 14, 14, 512) 2359808

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block5\_conv2 (Conv2D) (None, 14, 14, 512) 2359808

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block5\_conv3 (Conv2D) (None, 14, 14, 512) 2359808

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

block5\_pool (MaxPooling2D) (None, 7, 7, 512) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

flatten (Flatten) (None, 25088) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

fc1 (Dense) (None, 4096) 102764544

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

fc2 (Dense) (None, 4096) 16781312

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

predictions (Dense) (None, 1000) 4097000

=================================================================

Total params: 138,357,544

Trainable params: 138,357,544

Non-trainable params: 0

Tabela ‑ Parametrii VGG16

Modelul e capabil sa clasifice imaginile în 1000 de clase după cum putem observa din ultimul strat. Pe noi ne interesează stratul **fc2** ce produce 4096 de valori pentru imaginea de la intrare. Pentru a îmbunătății timpul de antrenare am spart modelul in 2, modelul convoluțional (encoderul) și modelul recurent(decoderul). Vom trece pozele din seturile din antrenare și validare prin primul model, VGG16, și vom salva cele 4096 de valori corespondente fiecare poze într-un fișier .pkl, pentru a le putea ulterior încărca când vom antrena rețeaua recurentă. Cu toate acestea numărul de stări interne ale rețelei convoluționale este de 512, așa ca vom avea nevoie de un strat dens care sa reducă dimensiunea de la 4092 la 512 intrări.

Pasul următor este să convertim cuvintele folosite pentru a descrie conținutul imaginilor din setul de antrenare și validare în numere întregi. Acest proces este cunoscut ca “tokenizare”, este utilizat in algoritmii de ML care lucrează pe text, deoarece rețelele neuronale nu pot opera direct pe cuvinte. Valorile întregi sunt convertite în vectori de lungimi egale ale căror elemente au valori raționale(floating). Acest va reprezenta stratul de embedding. Începutul și finalul propoziției/descrierii vor fi marcate prin 2 puțin probabile sa fac sens in contextul descrierii conținutului din imagine: 'ssss ' pentru start și ‘eeee’ pentru finalul ei.

Astfel, descrierile unei imagini înainte de tokenizare vor arata astfel:

['ssss Closeup of bins of food that include broccoli and bread. eeee',

'ssss A meal is presented in brightly colored plastic trays. eeee',

'ssss there are containers filled with different kinds of foods eeee',

'ssss Colorful dishes holding meat, vegetables, fruit, and bread. eeee',

'ssss A bunch of trays that have different food. eeee']

și post tokenizare:

[[2, 841, 5, 2864, 5, 61, 26, 1984, 238, 9, 433, 3],

[2, 1, 429, 10, 3310, 7, 1025, 390, 501, 1110, 3],

[2, 63, 19, 993, 143, 8, 190, 958, 5, 743, 3],

[2, 299, 725, 25, 343, 208, 264, 9, 433, 3],

[2, 1, 170, 5, 1110, 26, 446, 190, 61, 3]]

unde 2 și 3 reprezintă indexul pentru marker-ul de început, respectiv sfârșit.

Fiecare imagine are cel puțin 5 descrieri. Rețeaua recurentă va primi la intrare datele salvate anterior în fișierul .pkl, ce reprezintă ieșirea penultimului strat din VGG16. Numim epoca, trecerea unui număr de poze egal cu numărul de poze din setul de antrenare prin modelul nostru si ajustarea ponderilor pentru fiecare transă din acest set. Pentru a face antrenarea cât mai obiectiva pentru construirea unei tranșe de antrenare se va lua o poza la întâmplare din setul de antrenare alături de o descriere la întâmplare ce îi revine. Detrimentul acestei tehnici este ca o poza poate trece prin model de mai multe ori prin model in procesul de antrenare, în timp ce altele pot să nu se regăsească deloc în setul de antrenare

Am importat din tensorflow.keras.layers o implementare a GRU-ului și am construit modelul în felul următor.

* Din cele 4098 de valori obținute din rețeaua convoluțională cream o secvența de 128 de vectori prin embedding

decoder\_embedding=Embedding(input\_dim=num\_words,output\_dim=embedding\_size, name='decoder\_embedding')

* Primind datele de la stratul de embedding, acest strat aduce datele de intrare la în formatul unui vector de 512 elemente

state\_size=512  
decoder\_transfer\_map=Dense(state\_size,activation='tanh',name='decoder\_transfer\_map')

* Am aplicat o funcție de activare *tanh* pentru a obține valori cuprinse între -1 și 1

decoder\_transfer\_map=Dense(state\_size,activation='tanh',name='decoder\_transfer\_map')

* Cream 3 starturi recurente.

decoder\_gru1 = GRU(state\_size, name='decoder\_gru1', return\_sequences=True)

decoder\_gru2 = GRU(state\_size, name='decoder\_gru2', return\_sequences=True)

decoder\_gru3 = GRU(state\_size, name='decoder\_gru3', return\_sequences=True)

net = decoder\_gru1(net, initial\_state=initial\_state)

net = decoder\_gru2(net, initial\_state=initial\_state)

net = decoder\_gru3(net, initial\_state=initial\_state)

* Ultimul strat va fi reprezentat de un strat dens care va face o codare de tipul ‘one-hot’ pentru ieșirea rețelei recurente

decoder\_dense = Dense(num\_words,activation='softmax',name='decoder\_output')

decoder\_output = decoder\_dense(net)

Modelul arată astfel:

Model: "model\_1"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param

===============================================================

decoder\_input (InputLayer) [(None, None)] 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

transfer\_values\_input (InputLay [(None, 4096)] 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

decoder\_embedding (Embedding) (None, None, 128) 1280000

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

decoder\_transfer\_map (Dense) (None, 512) 2097664

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

decoder\_gru1 (GRU) (None, None, 512) 986112

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

decoder\_gru2 (GRU) (None, None, 512) 1575936

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

decoder\_gru3 (GRU) (None, None, 512) 1575936

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

decoder\_output (Dense) (None, None, 10000) 5130000

===============================================================

Total params: 12,645,648

Trainable params: 12,645,648

Tabela ‑ Model rețea recurentă

### Antrenarea

În încercarea obținerii unor rezultate cât mai bune am încercat să variem diferiți “hyperparametri” pentru a îmbunătății acuratețea modelului nostru. Am aplicat mai multe variații asupra procesului de învățare asupra, ratei de învățare, optimizatorului, dimensiunii lotului și nu in ultimul rând, modelului recurent. Mai jos sunt enumerate câteva din încercările de îmbunătățire cât și soluția cu care s-a decis să se meargă mai departe.

1. În prima încercare am utilizat ADAM ca funcție de cost cu o rată fixă de învățare 10e-3, timp de 20 e epoci. Am observat că în primele epoci funcția de cost a scăzut constant pana la 1. Din epoca a 10-a valoarea funcției de cost a început sa fluctueze fără a mai converge către un minim
2. În încercarea îmbunătățirii am schimbat funcția de cost în RMSprop și am observat că deși mai încet decât cu configurația de hiper parametri anterior menționați, modelul a continuat sa conveargă către un minim local
3. Pentru a micșora funcția am încercat să modific modelul crescând numărul de neuroni din stratul de transfer de la 512 la 1024. Acest lucru a îngreunat antrenarea, timpul per epoca crescând de 4 ori. Nici rezultatele nu au fost vizibil mai bune
4. Combinația de parametrii care a dat cele mai bune rezultate a avut un optimizator RMSprop cu o rată de învățare variabilă, timp de 20 de epoci, cu un o dimensiune a lotului de învățare de 3000 de imagini. O epoca a durat în acest caz aproximativ 5 ore. Mai jos se poate observa evoluția funcției de cost.

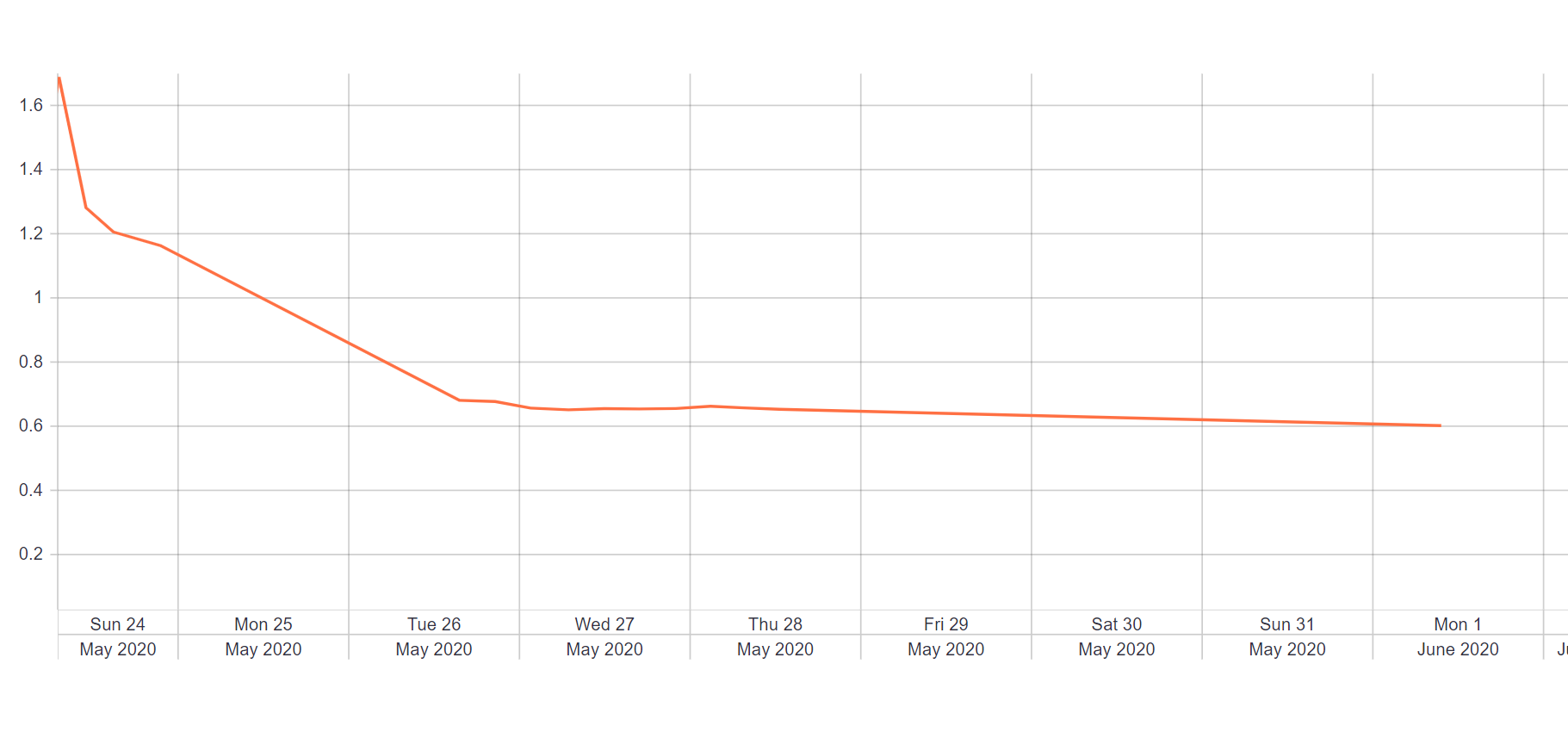


Figura ‑ Evoluția funcției de cost

## Rezultate

Pe ultima soluție menționată anterior vom face o analiza a rezultatelor obținute de modelul nostru asupra unor imagini noi, cât și asupra unor imagini din setul de antrenare.

Pentru date noi:

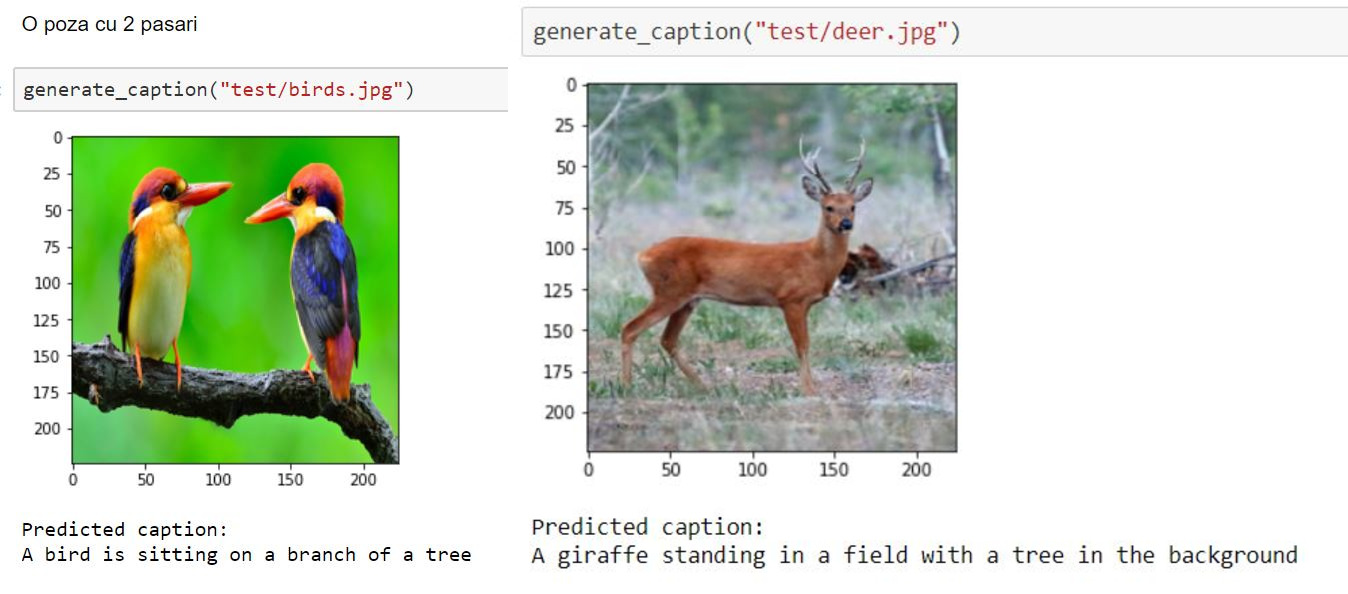


Figura ‑ Rezultate: Păsări și capra sălbatică

Se poate observa că modelul poate recunoaște că în imagine se află păsări și ca stau pe ramura unui copac însă nu și faptul că sunt 2 păsări. Datorită culorilor similare, modelul confunda capra sălbatică cu o girafă.



Figura ‑ Rezultate: Atlet și plajă

Observăm din această poză, că modelul e concentrat mai mult pe ceea ce se întâmpla în fundalul imaginii și nu în prim-planul ei. În a doua poză identifică cu succes mediul însă datorită amplasării destul de răsfirate a persoanelor și a dimensiunii lor reduse trage o concluzie incorectă asupra acțiunii de pe plajă.

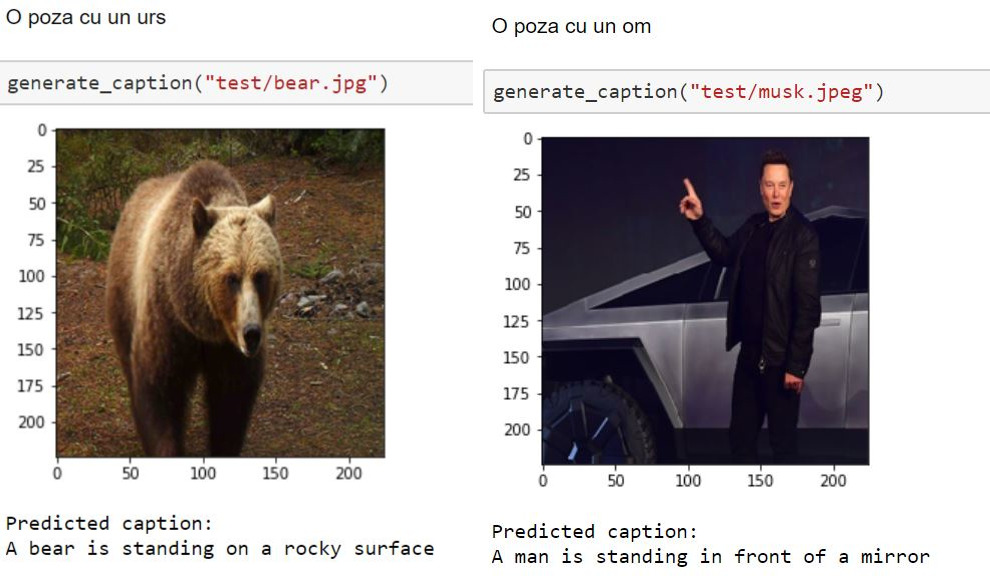


Figura ‑ Rezultate: Urs și Bărbat

Modelul este în a doua imagine aici păcălit de culoarea argintie a mașinii, și identifică bărbatul ca fiind în fața unei oglinzi.

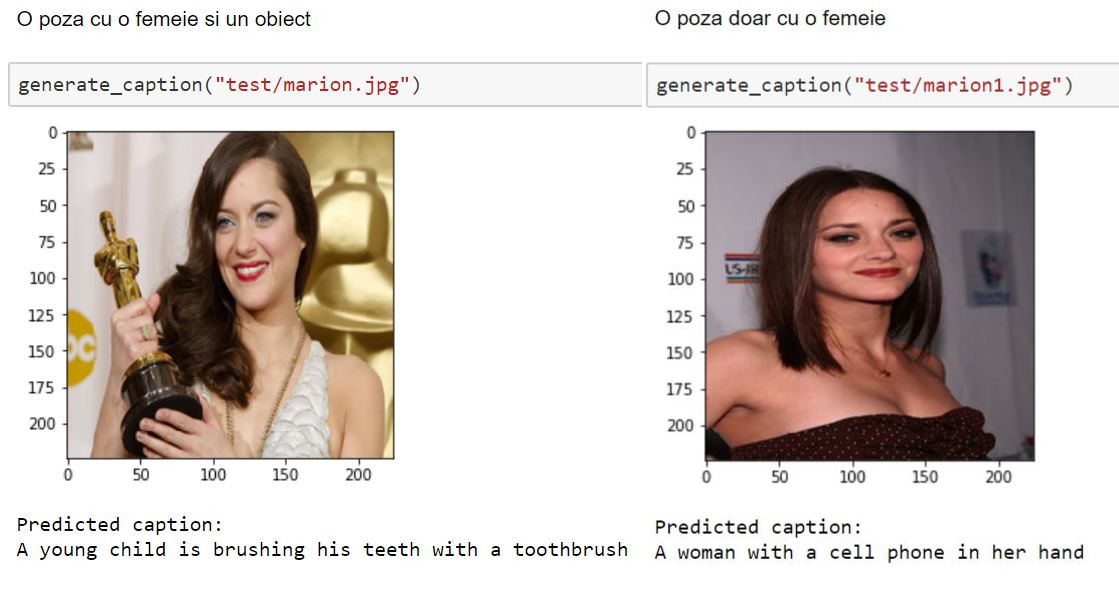


Figura ‑ Rezultate: Femeie în diferite contexte

Modelul pare să aibă o problemă în a identifica femei atunci când poza nu este una de portret și chiar și atunci acesta v-a returna cu preponderența ultima descriere.

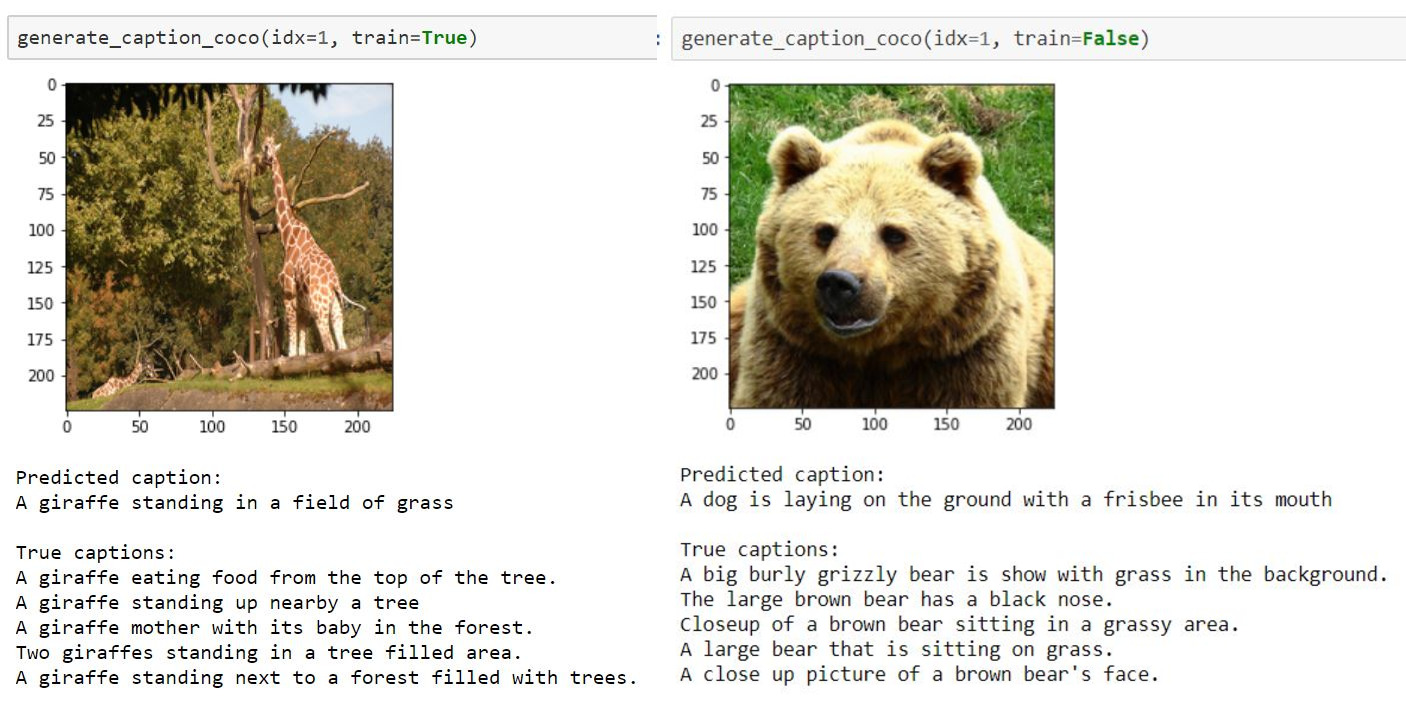


Figura ‑ Rezultate: Date din setul de antrenare

Pentru scena în care este prezentată o girafă modelul se descurcă și este capabil să proceseze și informația de fundal similar cu descrierile din setul de antrenare. În cea de a doua poză unde informația este mai puțina, poza fiind in majoritate cu fața ursului, modelul identifică greșit animalul și adaugă informații eronate despre acțiunea din imagine.

# Aplicația Android

## Introducere

Pentru a demonstra utilitatea unei astfel de algoritm, precum descrierea conținutului vizual, am decis să creez o aplicație care să semene cu un produs pe care l-am putea găsi pe piață. Mediul cel mai la îndemână s-a dovedit a fi prin intermediul unei aplicații pentru terminalul mobil. În momentul procesarea și analiza imaginilor reprezintă o categorie de interes pentru dezvoltatorii de Android/IOS.

Android este un proiect open-source, și nu necesită nici un tip de hardware proprietar dezvoltatorilor pentru a scrie aplicații, așa cum este în cazul IOS. Răspândirea largă a aplicațiilor android pe telefoane, tablete, ceasuri, televizoare, cât și comunitatea activă de dezvoltatori au influențat decizia de a dezvolta un soft care să se folosească de modelul prezentat anterior pentru a genera descrieri pentru imaginilor utilizatorilor.

## Tehnologii folosite

Pentru a dezvolta aplicația am utilizat Android Studio pentru dezvoltarea aplicației, utilizând Kotlin ca principalul limbaj de dezvoltare. Spre deosebire de Java, codul este mult mai compact iar tratarea excepțiilor se face într-o manieră mai modernă. Principalul dezavantaj pe care Kotlin îl are este datorat vârstei sale, majoritatea aplicațiilor fiind scrie încă in Java.

Pentru definirea interfeței vizuale Android Studio folosește cod XML, într-o manieră similara cu html-ul. Mediul de dezvoltare îi permite utilizatorului să dezvolte atât utilizând limbajul scris, cât și cu ajutorul unui meniu drag&drop. Adesea lucrând la interfața vizuala am apelat la ambele modalități pentru a obține o experiența vizuală cât mai plăcută pentru utilizator. Butoanele vizibile în aplicație au fost create cu ajutorul unei aplicații web numite “Call-to-Action Button Optimizer”. Aplicația permite utilizatorului crearea unor butoane care să respect o paletă de culori similară. Logo Maker este o altă aplicație web care mi-a permis să creez un logo pentru softul nostru. Site-ul permite generarea de logo-uri sub forma de imagini vectorizate pentru a putea fi folosite atât în logoul aplicației cât și în layoutul ei.

### Android Studio

Android Studio este un mediu de dezvoltare (software development environment, sau integrated development environment - "mediu integrat de dezvoltare) pentru colaborarea cu platforma Android, anunțată pe data de 16 mai 2013 în cadrul conferinței I / O Google.

IDE-ul este disponibil gratuit începând cu versiunea 0.1, publicată în mai 2013, apoi a trecut la testarea beta, începând cu versiunea 0.8, care a fost lansată în iunie 2014. Prima versiune stabilă 1.0 a fost lansată în decembrie 2014, apoi suportul pentru pluginul Android Development Tools (ADT) pentru Eclipse a încetat.

Android Studio este bazat pe software-ul IntelliJ IDEA de la JetBrains, este instrumentul oficial de dezvoltare a aplicațiilor Android. Acest mediu de dezvoltare este disponibil pentru Windows, OS X și Linux. Pe 17 mai 2017, la conferința anuală Google I / O, Google a anunțat asistență pentru limbajul Kotlin utilizat de Android Studio ca limbaj de programare oficial pentru platforma Android, pe lângă Java și C ++. [16]

Android Studio permite utilizatorilor să creeze dispozitive virtuale android(AVD). Un dispozitiv virtual Android (AVD) este o configurație care definește caracteristicile unui telefon Android, tabletă, Android TV pe care dorim să le simulați în emulatorul Android. AVD Manager este o interfață în Android Studio care ne ajută să cream și să gestionăm AVD-uri.

### Dependinte(Gradle File)

În scopul ușurării dezvoltării am utilizat o serie de pachete/librarii externe pentru a exploata resursele hardware ale terminalului mobil

## Structura(Diagrama bloc explicata)

### Activitatea introductiva

### Activitatea de Log In

### Activitatea de Register

### Activitatea de Alegere a pozei

### Activitatea de obtinere a descrierei textuale

# Firebase

## Motivatie/Descriere

### Ce este Firebase

## Autentificare

## Docker?

## Hosting?/Flask

# Testare

## Descrierea Utilizarii

## Rezultate

## Scalabilitate

## Imbunatatiri

# Concluzie

# Bibliografie

|  |  |
| --- | --- |
| [1] | Google, „Androi Development,” Google, [Interactiv]. Available: https://developer.android.com/. [Accesat 18 1 2020]. |
| [2] | Android Authority, „Kotlin vs Java,” 18.10.2019. [Interactiv]. Available: https://www.androidauthority.com/kotlin-vs-java-783187/. [Accesat 18 1 2020]. |
| [3] | „Flask official page,” [Interactiv]. Available: https://flask.palletsprojects.com/en/1.1.x/. |
| [4] | „Rețele neuronale în TensorFlow,” [Interactiv]. Available: http://myac.xhost.ro/inva/Lab9/InvA\_Lab9.pdf. [Accesat 18 1 2020]. |
| [5] | HVASS-Labs, „Image Captioning,” [Interactiv]. Available: https://github.com/Hvass-Labs/TensorFlow-Tutorials/blob/master/22\_Image\_Captioning.ipynb. |
| [6] | C. Florea, „Curs MLAV,” [Interactiv]. Available: http://www.master-taid.ro/Cursuri/MLAV\_files/07\_08\_MLAV\_ConvNets\_CF.pdf. |
| [7] | M. A. C., Introduction to Machine Learning with Python: A Guide for Data Scientists, Paperback, 2012. |
| [8] | F.-F. L. &. J. J. &. S. Yeung, „C231,” 2017. [Interactiv]. Available: http://cs231n.stanford.edu/slides/2017/. [Accesat 2020]. |
| [9] | S. Tejani. [Interactiv]. Available: https://shafeentejani.github.io/assets/images/pooling.gif. |
| [10] | J. J. S. Y. Fei Fei LI, „CS231,” Stanford, 2017. |
| [11] | J. J. S. Y. Fei Fei Li, „cs231n\_2017\_lecture10,” 4 5 2017. [Interactiv]. Available: http://cs231n.stanford.edu/slides/2017/cs231n\_2017\_lecture10.pdf. |
| [12] | B. v. M. C. G. B. B. H. S. B. Kyunghyun Cho, „Learning Phrase Representations using RNN Encoder–Decoder,” [Interactiv]. Available: https://arxiv.org/pdf/1406.1078.pdf. |
| [13] | „Gated recurrent unit,” Wikipedia, [Interactiv]. Available: https://en.wikipedia.org/wiki/Gated\_recurrent\_unit#/media/File:Gated\_Recurrent\_Unit,\_base\_type.svg. |
| [14] | C. Florea, „Curs MLAV,” [Interactiv]. Available: http://www.master-taid.ro/Cursuri/MLAV\_files/11\_12\_MLAV\_En\_Transfer\_2018.pdf. |
| [15] | „Extract Features, Visualize Filters and Feature Maps in VGG16 and VGG19 CNN Models,” 2020. [Interactiv]. Available: https://mc.ai/extract-features-visualize-filters-and-feature-maps-in-vgg16-and-vgg19-cnn-models/. |