SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Klasifikacija videa

Antonio Borac
Voditelj: *Siniša Šegvić*

Zagreb, lipanj 2019.

SADRŽAJ

1. Uvod

U zadnjih nekoliko godina razvojem algoritama učenja i razvojem hardvera koji podržava izrazito paralelnu obradu podataka, ostvaren je ogroman napredak na području dubokog učenja. Razvijeni su konvolucijski modeli koji pronalaze primjenu na raznim problemima poput klasifikacije slika, klasifikacije videa i analize teksta. Takvi modeli na problemu klasifikacije slika danas nadmašuju čak i čovjeka.

Svaki dan se na popularne internet stranice postavljaju ogromne količine video sadržaja. Teško je ručno kontrolirati što se točno postavlja na takve stranice, pa ima smisla pokušati filtrirati takav video materijal s obzirom na prikladnost sadržaja za razne društvene kategorije. Tako bi mogli razvrstati sadržaj u kategorije poput sadržaja namijenjenog za djecu, sportskog sadržaja, znanstveno-popularnog sadržaja i slično. Za problem klasifikacije videa razvijeni su mnogobrojni modeli koji koriste reprezentaciju "vreće riječi" i klasifikatore poput stroja potpornih vektora (SVM)[6]. Ubrzanim razvojem konvolucijskih modela ostvaruju se dobri rezultati koji nadmašuju rezultate ostvarene takvim modelima. Tako Karpathy [6] razvija prvi konvolucijski model koji nadmašuje rezultate prethodnih modela na najpoznatijim skupovima za mjerenje performanse klasifikacije videa.

Klasifikacija videa ima velik potencijal primjene. Modele bi mogli iskoristiti u zdravstvu, sigurnosti, sportu i mnogim drugim područjima. Osim klasifikacije, razvijaju se modeli za razne druge probleme poput segmentacije videa, procjene dubine i procjene poze[3].

U okviru ovog rada nastojat ćemo prikazati tehnike koje se koriste za razvoj modela za video i rezultate koji su ostvareni korištenjem najuspješnijih modela na trenutno popularnim skupovima za klasifikaciju videa. Opisat ćemo popularne skupove podataka koji se koriste za treniranje modela za klasifikaciju videa. Analizirat ćemo što je video i kako se može upotrijebiti za klasifikaciju.

2. Što je video?

Na video možemo gledati kao niz slika u vremenu. Možemo reći da video predstavlja trodimenzionalan podatak - ima dvije prostorne dimenzije i jednu vremensku dimenziju. Jedan video bez kompresije može zauzimati mnogo prostora. Tako video vremenskog trajanja 10 sekundi snimljen kamerom koja snima 30 okvira u sekundi i koja snima pojedine okvire dimenzija 200x200 piksela, bez kompresije zauzima 1,152 MB. Jedna fotografija istih dimenzija zauzima 3.84 MB. Ako promotrimo pojedine odsječke videa okvir po okvir lako možemo primijetiti da su mnogi okviri jako slični - često će objekti na nekom odsječku videa biti statični ili će se tek malo pomaknuti između pojedinih okvira. Za treniranje dubokih modela nepotrebno je analizirati svaki pojedini okvir jer obično želimo uhvatiti kontekst cijelog videa za koji nam oni nisu bitni. Štoviše, obrada svakog pojedinačnog okvira bi bila nepotrebno zahtjevna. Za izgradnju naših modela najčešće ćemo posegnuti za uzorkovanjem tako da uzmemo samo nekoliko okvira po sekundi - ponekad će nam biti dovoljan i samo jedan okvir po sekundi, ovisno o videu kojeg uzorkujemo. Međutim, kako ne bi izgubili informaciju o pokretu između bliskih okvira koji može doprinijeti kvaliteti klasifikacije, možemo iskoristiti optički tok [4]. Optički tok, prema definiciji, predstavlja prividno kretanje svjetlosnog uzorka između okvira. Osim vizualnih komponenti, video obično sadrži i zvuk koji također može pomoći pri klasifikaciji video sadržaja što bi svakako bilo zanimljivo razmotriti.

3. Skupovi podataka za klasifikaciju videa

3.1. UCF-101

Danas postoji niz standardnih skupova podataka koji se koriste za provjeru razvijenih modela. UCF101 [9] je skup podataka koji sadrži kratke video isječke različitih ljudskih aktivnosti. Sastoji se od 101 kategorije raznovrsnih aktivnosti poput raftinga, šminkanja i sviranja raznih instrumenata. Skup se sastoji od preko 13000 snimaka koji

Slika 3.1: 6 klasa iz skupa UCF101. Slika je preuzeta iz [9].

ukupno traju više od 27 sati. Zbog velikog broja klasa i relativno velike varijacije među klasama, ovaj se skup smatra dobrim pokazateljem performanse modela. Problem s podacima iz ovog skupa je što su videozapisi unutar neke klase snimani pod sličnim osvjetljenjem i pod sličnim kutom kamere pa je varijacija unutar neke klase relativno mala. To ne odražava stvarne videe koji za neku akciju mogu biti dosta različiti.

3.2. YouTube-8m

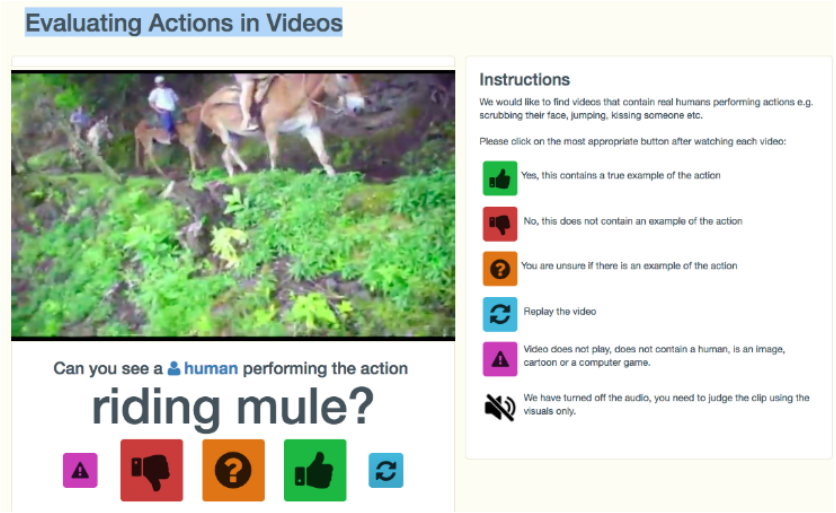
Skup YouTube-8m [2] je skup od preko 8 milijuna video snimaka koji ukupno traju oko 500 000 sati. Skup je označen s 4803 klase pri čemu pojedini podatak može biti označen s više od jedne klase. Za razliku od skupa UCF101, podaci iz ovog skupa bolje

Slika 3.2: Slika predstavlja 200 najčešćih oznaka u YouTube-8m skupu. Veličina fonta odražava koliko često se oznaka pojavljuje u skupu. Slika je preuzeta iz [2].

karakteriziraju videozapise kakve obično možemo pronaći jer je sastavljen od stvarnih YouTube videa označenih pomoću YouTube anotacijskog sustava koji označava glavne teme unutar videa.

3.3. Kinetics Human Action Video Detection

Kinetics Human Action Video Detection je relativno nov skup podataka iz 2017. godine koji je za dva reda veličine veći od skupova poput UCf-101[3] - ukupno se sastoji od preko 300 000 podataka. Skup se sastoji od 400 razreda ljudskih akcija od kojih svaka akcija ima preko 400 videa. Svaki video je dobiven kao kratki isječak iz YouTube videa [7]. Prednost skupa u odnosu na postojeće skupove je što je svaki isječak dobiven kao isječak iz različitog originalnog videa za razliku od skupa poput UCF-101 gdje je iz jednog videa izvađeno i do 7 isječaka. Tako skupljeni podaci imaju mnogo veću varijancu što pogoduje treniranju i testiranju dubokih modela za klasifikaciju videa. Također, svi videi su snimani u neprofesionalnim uvjetima slično kao i podaci u skupu YouTube-8m. Skup je ručno označen oznakama pomoću Amazon Mehanical Turkers tehnologije. Akcija svakog od isječaka je trebala biti potvrđena tri od pet puta da bi snimak bio prihvaćen.



Slika 3.3: Postupak označavanja videa pri izradi skupa Kinetics pomoću Amazon Mechanical Turk. Slika je preuzeta iz [7].

Za očekivati je da će razvojem modela za klasifikaciju videa skup postepeno preuzimati ulogu UCF-101 skupa kao osnovni skup za evaluaciju razvijenih modela.

4. Modeli za klasifikaciju videa

Svrha ovog poglavlja je opisati tehnike koje se koriste za izgradnju modela za klasifikaciju videa. Postoji mnogo različitih, više ili manje uspješnih pristupa ovom problemu. Pokušat ćemo dati objašnjenje koliko su pojedini pristupi dobri za konstrukciju modela za klasifikaciju videa. U sljedećem poglavlju ćemo opisati pojedine arhitekture nastale kombinacijom ovih tehnika.

4.1. Naivan pristup konvolucijskim modelima

Ovaj pristup iskorištava strukturu videa kao podatka. Kako smo ranije naveli, video je niz nanizanih okvira u vremenu. Svaki od tih okvira možemo promatrati kao jednu statičnu sliku koju nastojimo klasificirati u neki razred. Tako svaki okvir u videu predstavlja klasifikacijski problem za sebe. Nakon što smo izračunali predikcije svakog odvojenog okvira, izračunamo konačnu predikciju kao prosječnu predikciju svih okvira. Ovakav pristup bi loše radio iz nekoliko razloga. Naime, iako je video niz okvira u vremenu, ne možemo iz svakog okvira jasno razlučiti semantiku. Pojedini okvir ne mora sadržavati objekte koji su nam bitni za klasifikaciju videa u definirane razrede - okvir može prikazivati pod koji nam potencijalno ne koristi u predikciji. Na ovaj način svaki okvir može klasificirati na temelju onoga što se nalazi u tom okviru ne koristeći informaciju što je prikazano na ostalim okvirima [8].

Ovaj pristup ne može eksplicitno iskoristiti informaciju o pokretu objekta između pojedinih okvira što može imati značajan utjecaj na problemima poput klasifikacije ljudskih radnji. Zaključak ovog pokušaja je da modeli koje gradimo ne mogu video promatrati samo kao niz okvira - pokazat će se da temporalna komponenta videa ima značajan utjecaj na klasifikaciju.

4.2. Arhitekture sažimajućih značajki

Mana prethodne arhitekture je što pri klasifikaciji nije uzimala u obzir međusoban utjecaj pojedinih okvira. Taj problem je moguće riješiti uvođenjem slojeva sažimanja koji omogućavaju kombinaciju dohvaćenih značajki iz pojedinih okvira. Ideja je da se za svaki okvir konstruira konvolucijski model te da se na različite načine uz pomoć slojeva sažimanja kombiniraju izlazi svih okvira. Sažete značajke je moguće kombinirati potpuno povezanim slojevima i na izlazu dati predikciju kojem razredu video pripada. Prednost takvih arhitektura je što omogućuje klasifikaciju na temelju cijele sekvence pojedinog videa. Pri sažimanju se najčešće koristi sažimanje maksimalnom vrijednosti jer alternative poput prosječnog sažimanja i potpuno povezanog sloja unose previše gradijenata te usporavaju treniranje.

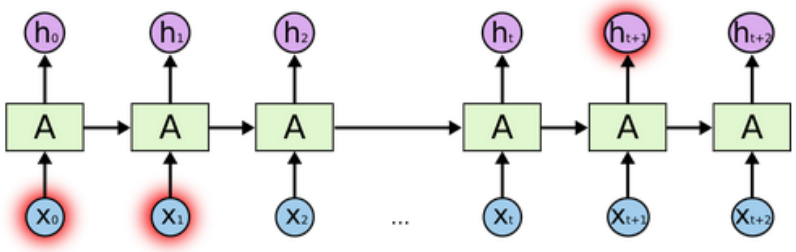
4.3. Long Short Term Memory (LSTM)

LSTM-ovi su specijalizacija povratnih neuronskih mreža. Za potrebe shvaćanja načina na koji ovi modeli rade, u nastavku dajemo uvod u način rada povratnih neuronskih mreža.

4.3.1. Povratne neruonske mreže

Ideja iza povratnih neuronskih mreža je iskoristiti sekvencijsku strukturu podataka. Klasični modeli bazirani na neuronskim mrežama pretpostavljaju da su podaci međusobno neovisni. Povratne neuronske mreže pretpostavljaju da to nije slučaj. Primjer problema s međusobno ovisnim podacima je prevođenje rečenica iz jednog jezika u drugi. Svaka riječ u rečenici ovisi o ostalim riječima u rečenici. To možemo lako uočiti ako pokušamo predvidjeti koja riječ dolazi na kraju sljedeće rečenice: "Oblaci su na ?.". Prilično je jasno da rečenica treba završiti sa "nebu".

Povratne neuronske mreže zovemo povratnim jer prilikom obrade trenutnog podatka koriste rezultate obrade prethodnih podataka u nekoj sekvenci.



Slika 4.1: Prikaz osnovne povratne neuronske mreže. Slika je preuzeta iz [1].

Za ulaznu sekvencu $x = (x_1, ..., x_T)$ standardna povratna neuronska mreža računa sekvencu skrivenog sloja $h = (h_1, ..., h_T)$ i izlaznu sekvencu $y = (y_1, ..., y_T)$. Za $t = 1$ do T izlaze računamo prema sljedećim formulama:

$$h_t = \mathcal{H}(W_{ih}x_t + W_{hh}h_{t-1} + b_h) \tag{4.1}$$

$$y_t = W_{ho}h_t + b_o. \tag{4.2}$$

Formula (4.1) računa izlaze skrivenog sloja za svaki ulaz sekvence. Funkcija $\mathcal{H}()$ predstavlja nelinearnu funkciju poput funkcije $sigmoid$ ili $tanh$. Matrica W_{ih} je matrica težina između ulaznog i skrivenog neurona, matrica W_{hh} je matrica težina između dva skrivena neurona i matrica W_{ho} je matrica težina između izlaza i skrivenog neurona. Vektori b_h i b_o predstavljaju pomak za računanje vrijednosti aktivacije skrivenog i izlaznog sloja. Bitno je napomenuti da su svi navedeni parametri jednaki kod računanja različitih ulaza sekvence.

Navedene formule i prikazana slika objašnjavaju osnovni koncept iza povratne neuronske mreže. Sada je moguće informaciju iz prošlosti koristiti za izračun trenutnog izlaza. Ovo svojstvo povratnih neuronskih mreža uvelike pomaže pri obradi videa. Način rada povratne neuronske mreže nije toliko neintuitivan koliko se potencijalno čini na prvu. Na odluke čovjeka tijekom razmišljanja ne utječu samo podaci iz njegove okoline već i prethodna iskustva što na neki način sliči radu povratne neuronske mreže. Međutim, ispostavlja se da obične povratne neuronske mreže imaju problem zbog kojeg ne pokazuju dobre rezultate na problemu klasifikacije videa. Slabo detektiraju dugoročne vremenske odnose pojedinih okvira. Zbog toga su dizajnirane LSTM ćelije koje popravljaju taj problem.

4.3.2. LSTM

LSTM za razliku od povratnih neuronskih mreža koristi memorijske ćelije za pohranu i propagaciju informacija što pomaže pri otkrivanju dugoročnih vremenskih odnosa.

Razlika u izračunu u odnosu na obične povratne neuronske mreže je pri izračunu izlaza skrivenog sloja. Izlaz h_t računamo na sljedeći način:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{4.3}$$

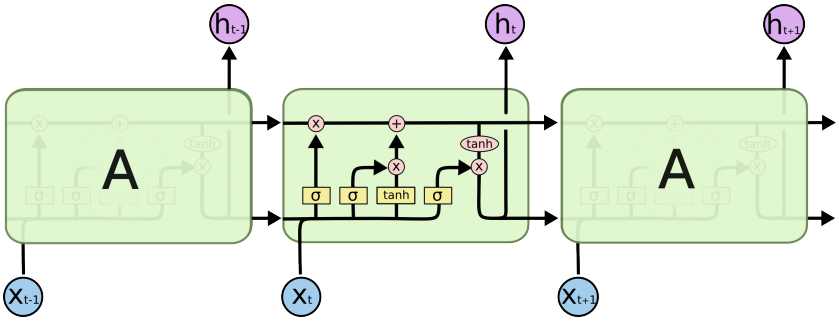
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4.4}$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{4.5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{4.6}$$

$$h_t = o_t tanh(c_t) \tag{4.7}$$

gdje je σ sigmoidalna funkcija. i, f i o nazivamo redom $ulazna\ vrata$, $vrata\ zaboravljanja$ i $izlazna\ vrata$. c nazivamo LSTM ćelijom.



Slika 4.2: Prikaz LSTM ćelije. Slika je preuzeta iz [1].

Svaka ćelija čuva jednu realnu vrijednost koju mreža može potisnuti faktorom f ili aditivno pojačati faktorom i. Te ćelije omogućavaju mreži da pamti vrijednosti kroz vrijeme što omogućava detekciju dugoročnih ovisnosti između podataka. Povratne neuronske mreže koriste derivabilne funkcije pa je postupak treniranja mreže moguće ostvariti unatražnom propagacijom. Za ove mreže je specifično što će prilikom propagacije greške unatrag gradijenti potrebni za izračun daljnjih gradijenata u određenom trenutku dolaziti iz izlaza mreže i budućeg trenutka.

4.4. Optički tok

Optički tok je jedna od najbitnijih tehnika u području računalnog vida [10]. Pronalazi primjenu u mnogim područjima poput prepoznavanja akcija, analize videa i autonomne vožnje. Generalno, optički tok nastoji procijeniti pomak svjetlosnih uzoraka kroz scenu u stvarnom vremenu. Taj pomak aproksimira kretnju objekata između dva

susjedna okvira u videu. Ova informacija može mnogo pomoći pri klasifikaciji videa koji prikazuju primjerice ljudske radnje.

(a) First frame (b) Second frame (c) Optical flow field

Slika 4.3: Prve dvije slike prikazuju dva susjedna okvira. Treća slika prikazuje optički tok između slika. Tok je vizualiziran na način da boja predstavlja intenzitet i smjer pomaka objekta u videu. Slike preuzete iz [12].

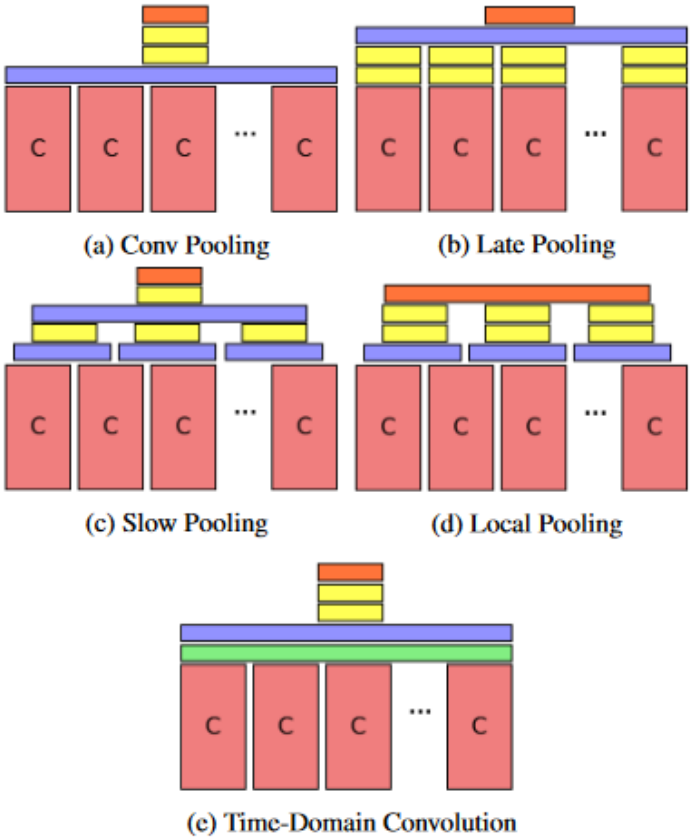
Procjena optičkog toka je problem za sebe, pojedini autori [10] predlažu modele koji aproksimiraju optički tok pomoću konvolucijskih modela.

5. Arhitekture modela

U ovom poglavlju ćemo predstaviti nekoliko poznatih modela za klasifikaciju ljudskih radnji na poznatom skupu UCF-101, koji koriste neke od opisanih tehnika u prethodnom poglavlju. U sljedećem poglavlju ćemo sažeti ostvarene rezultate i usporediti performanse modela.

5.1. Arhitekture sažimajućih značajki

Rad [8] ispituje ponašanje nekoliko različitih arhitektura koje se baziraju na slojevima sažimanja. Koriste se sažimanje maksimumom zbog manje količine gradijenata. Svim arhitekturama je zajedničko da ulazni video uzorkuju na način da uzmu određen broj uzoraka iz videa. Svaki uzorak propuštaju kroz konvolucijsku mrežu. Bitno je napomenuti da predložen model dijeli parametre između okvira. Autori izlaze iz konvolucijskih modela kombiniraju na različite načine koje možemo vidjeti na slici 5.1 nastojeći pronaći koji model najbolje radi.

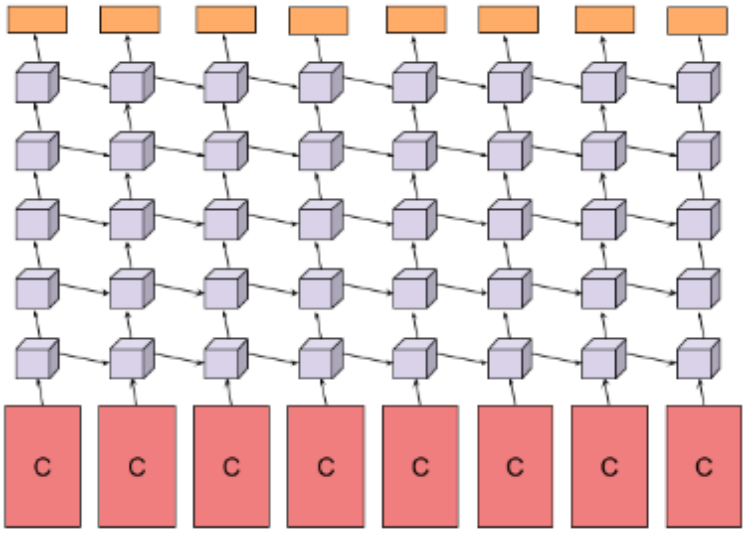


Slika 5.1: Prikaz raznih arhitektura sažimajućih značajki. Slike preuzete iz [8].

Na slici 5.1 $a)$ se radi sažimanje maksimumom, nakon kojeg dolaze dva potpuno povezana sloja čiji se izlaz propušta kroz funkciju $softmax$. Na modelu prikazanom na 5.1$b)$ se svaki izlaz konvolucijskog modela dovodi na dva potpuno povezana sloja. Nad njihovim izlazom se obavlja sažimanje maksimumom te se izlaz ponovno dovodi na funkciju $softmax$. Na slici 5.1 $c)$ se obavlja sažimanje maksimumom izlaza konvolucijskog modela na dva susjedna podatka. Zatim se na svaki sloj sažimanja povezuje potpuno povezani sloj te se izlazi takvih slojeva ponovno sažimaju maksimumom. Dodaje se još jedan potpuno povezan sloj i funkcija $softmax$. Slika 5.1 $d)$ prikazuje sičnu ideju, razlika u odnosu na $b)$ je što se na svaki sloj sažimanja dodaju dva potpuno povezana sloja. Zatim se postavlja funkcija $softmax$ koja se računa na temelju izlaza svakog potpuno povezanog sloja. Prednost ovog modela je što čuva informaciju o poziciji značajki u vremenu za razliku od modela koji imaju dodatan sloj sažimanja. Razlike 5.1 $e)$ u odnosu na 5.1 $a)$ je što se na izlaz konvolucijskih slojeva dodaje još jedan konvolucijski sloj koji obavlja operaciju konvolucije nad 10 okvira.

5.2. LSTM arhitektura

U radu [8] je osim arhitekture sažimajućih slojeva ispitana i arhitektura koja se bazira na LSTM ćelijama prikazanim na slici 4.2. Složen je model od 5 nanizanih LSTM slojeva, svaki sa 512 ćelija. Slojevi su nanizani jedan na drugi na način da je izlaz jednog sloja ulaz u drugi sloj. Na prvi sloj se dovode izlazi iz niza konvolucijskih slojeva.



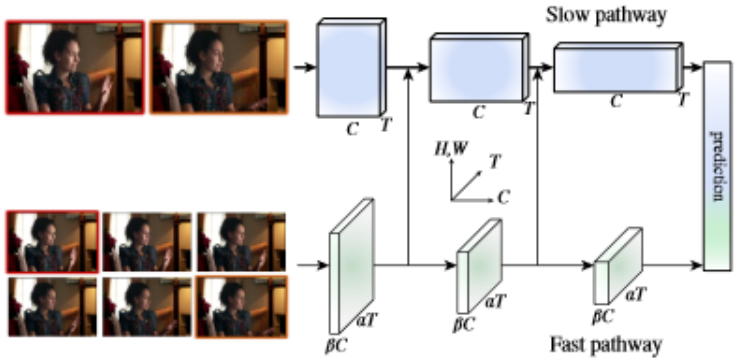
Slika 5.2: Prikaz predložene arhitekure bazirane na LSTM ćelijama. Slike preuzete iz [8].

Ovaj model predviđa razred modela u svakom vremenskom trenutku. Kao u prethodnoj arhitekturi, svi parametri su dijeljeni između vremenskih trenutaka [8]. Konačna predikcija se računa težinska linearna interpolacija tijekom vremena koja se sumira i vrati maksimalna predikcija.

5.3. Brza-Spora mreža za klasifikaciju videa

U radu [4] je predložena nešto drugačija arhitektura modela za klasifikaciju videa i detekciju objekta u videu. Autori pretpostavljaju da bi model ostvario bolji rezultat klasifikacije ako "hvata" prostorne strukture i vremenske događaje zasebno. Koriste dvije odvojene staze u modelu: brza staza i spora staza. Obje staze na ulaz dobivaju iste originalne podatke. Razlika je u tome što se na ulaz brze staze dovode podaci koji su uzorkovani često, a na ulaz spore staze podaci koji su uzorkovani rijetko u odnosu na brzinu uzorkovanja kod brze staze. U kontekstu prepoznavanja akcija i objekata, kategorična prostorna semantika sadržaja videa se sporo mijenja, dok se neka radnja

koja se vrši tokom videa mijenja često u odnosu na objekte koji radnju izvršavaju. Ovu ideju možemo predočiti sljedećim primjerom. Zamislimo snimak ruke koja maše. Ruka se tokom cijelog videa može klasificirati kao ruka bez obzira na trenutne uvjete u sceni poput položaja ruke ili količine svjetla. Radnja se za razliku od objekta može odvijati dosta brzo pa ju treba prikazati sa više okvira kako bi se jasno izvukao kontekst radnje.



Slika 5.3: Prikaz predložene Slow-Fast arhitekture. Uočiti odnos broja kanala i vremenske dimenzije između staza. Slike preuzete iz [4].

Za detekciju samog pokreta nam nije bitna toliko prostorna semantika koliko je bitna vremenska semantika. Zato kako bi zbog čestog uzorkovanja brza staza bila manje računalno zahtjevna, na ulaz se dovode podaci manjih prostornih dimenzija i sa manjim brojem kanala. Ako definiramo faktor τ kao broj svakih koliko okvira uzorkujemo, tada je broj okvira koji se dovodi na ulaz spore mreže oko $\tau x T$ gdje je T trajanje cijelog videa. U slučaju brze staze uzorkujemo svakih τ/α okvira, gdje $\alpha > 1$ predstavlja odnos brzine uzorkovanja brze staze naspram spore staze. Tako ako je faktor $\alpha = 8$, brza staza će uzorkovati 8 puta češće od spore staze. Također uvodimo faktor $\beta < 1$ kao odnos broja kanala brze staze naspram spore staze. U ovom konekstu opisane staze mogu biti konvolucijski modeli ili primjerice rezidualne arhitekture[1] kojima se nećemo baviti u okviru ovog izlaganja. Nakon obrade brze i spore staze, uvode se lateralne veze koje propagiraju informacije iz brze staze u sporu stazu. Lateralne veze provode transformaciju značajki kako bi se izjednačile dimenzije mape značajki brze i spore staze.

[1]Originalni rad sa Resnet modelima koji ostvaruje jako dobre rezultate na problemu klasifikacije slika https://arxiv.org/abs/1512.03385

6. Rezultati

U ovom poglavlju ćemo prikazati rezultate arhitektura koje smo opisali u ovom radu.

Tablica 6.1: Prikaz rezultata raznih Feature Pooling arhitektura na Sports-1M skupu.Tablica je preuzeta iz [8].

Metoda	top-1	top-5
$ConvPooling$	68.7	89.3
$LatePooling$	65.1	87.2
$SlowPooling$	67.1	88.4
$LocalPooling$	68.1	88.9
$Time - DomainConvolution$	64.2	87.2

Drugi stupac u tablici prikazuje postotak slučajeva u kojima je najveća vrijednost nakon zadnjeg softmaksa u modelu bila na poziciji točnog razreda. Treći stupac tablice prikazuje postotak slučajva u kojima je točna oznaka razreda bila u najvećih 5 vrijednosti izlaza modela. U tablici 6.1 vidimo da se najboljom pokazala arhitektura koja obavlja sažimanje odmah nakon konvolucijskih slojeva prikazana na slici 5.1 $a)$. Zbog toga su autori eksperimente nastavili sa $Conv\ Pooling$ arhitekturom.

Tablica 6.2: Prikaz rezultata LSTM i arhitektura sažimajućih značajki na Sports-1M skupu. Tablica je preuzeta iz [8].

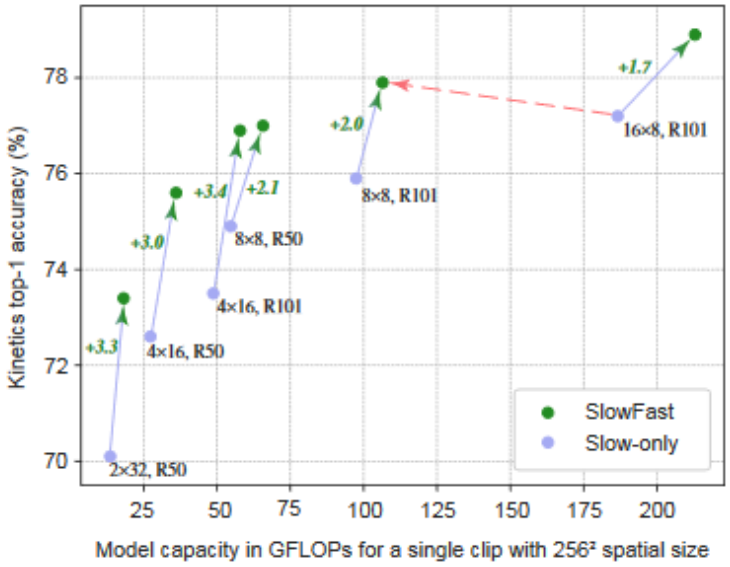
Metoda	top-1	top-5
LSTM sa optičkim tokom	59.7	81.4
LSTM sa izvornim okvirima	72.1	90.6
LSTM sa izvornim okvirima i optičkim tokom	73.1	90.5
Conv Pooling sa izvornim slikama	71.7	90.4
Conv Pooling sa izvornim slikama i optičkim tokom	71.8	90.4

Rezultati pokazuju da optički tok sam za sebe na ulazu nije dovoljan za problem klasifikacije. Najbolje ga je kombinirati sa izvornim slikama . Tako modeli bazirani na LSTM ćelijama i arhitekture sažimajućih značajki imaju bolju performansu sa kombinacijom izvornih slika i optičkim tokom.

Tablica 6.3: Prikaz najboljih rezultata SlowFast modela sa Resnet101 arhitekturom sa nelokalnim operacijama na nekoliko različitih skupova. Tablica je preuzeta iz [4].

Skup podataka	top-1	top-5
Kinetics-400	79.8	93.9
Kinetics-600	81.8	95.1

Model baziran na Resnet101 arhitekturi sa nelokalnim operacijama [11] postiže trenutno najbolje rezultate na gore prikazanim skupovima. Također, model sa brzim i sporim stazama konzistentno ostvaruje bolje rezultate od modela sa samo sporom stazom što možemo vidjeti na slici.



Slika 6.1: Prikaz rezultata SlowFast arhitekture u odnosu na Slow-only arhitekturu na modelima različite složenosti. Slike preuzete iz [4].

7. Zaključak

Klasifikacija videa je široko područje sa relativno širokom primjenom u praksi. Trenutno za probleme analize slika postoji mnogo više skupova i modela. Postoje pristupi koji nastoje iskoristiti modele za analizu slike u analizi videa. Tako rad [5] iskorištava modele za semantičku segmentaciju slika u problemu segmentacije videa. Razvojem modela za analizu videa i razvojem novih skupova podataka, rezultati će početi dostizati kvalitetu analize kakvu danas imamo na slikama. Tehnike predstavljene u ovom radu predstavljaju pristupe koji su se pokazali dobrim u praksi.Tako se arhitekture bazirane na LSTM-ovima ponašaju dobro na problemu klasifikacije videa. Optički tok u kombinaciji sa izvornim slikama se pokazao kao najbolji pristup u okviru rada [8] gdje je postignut rezultat od 73.1% točnosti na skupu za testiranje Sports-1M. Prednost arhitektura baziranih na LSTM-ovima je što možemo klasificirati video trajanja nekoliko minuta zahvaljući tome što LSTM može iskoristiti duge sekvence za klasifikaciju. Mana takvih arhitektura je što je prisutan problem nestajućih gradijenata prilikom učenja ali u znatno manjoj mjeri nego kod običnih povratnih arhitektura. Arhitektura bazirana na brzim i sporim stazama se pokazala obećavajućom na problemu detekcije i klasifikacije videa. Ostvaren je trenutno najbolji rezultat na skupovima AVA, Kinetics-400 i Kinetics-600. Problem te arhitekture je velika računalna složenost zbog koje je potrebno pažljivo birati parametre komponenti poput veličine rezidualnih modela koji se koriste. Koncept pokazan na toj arhitekturi daljnjim razvojem obećava veći uspjeh i na ostalim poznatim skupovima podataka.

8. Literatura

[1] Understanding LSTM Networks – colah's blog, May 2019. URL http://colah.github.io/posts/2015-08-Understanding-LSTMs. [Online; accessed 11. May 2019].

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, i Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. U *arXiv:1609.08675*, 2016. URL https://arxiv.org/pdf/1609.08675v1.pdf.

[3] João Carreira i Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. URL http://arxiv.org/abs/1705.07750.

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, i Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018. URL http://arxiv.org/abs/1812.03982.

[5] Raghudeep Gadde, Varun Jampani, i Peter V. Gehler. Semantic video cnns through representation warping. *CoRR*, abs/1708.03088, 2017. URL http://arxiv.org/abs/1708.03088.

[6] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, i Li Fei-Fei. Large-scale video classification with convolutional neural networks. U *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, i Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

[8] Joe Yue-Hei Ng, Matthew J. Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, i George Toderici. Beyond short snippets: Deep networks for video classification. *CoRR*, abs/1503.08909, 2015. URL http://arxiv.org/abs/1503.08909.

[9] Khurram Soomro, Amir Roshan Zamir, i Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

[10] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, i Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *CoRR*, abs/1709.02371, 2017. URL http://arxiv.org/abs/1709.02371.

[11] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, i Kaiming He. Non-local neural networks. *CoRR*, abs/1711.07971, 2017. URL http://arxiv.org/abs/1711.07971.

[12] C. Zach, T. Pock, i H. Bischof. A duality based approach for realtime tv-l1 optical flow. U *Proceedings of the 29th DAGM Conference on Pattern Recognition*, stranice 214–223, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-74933-2. URL http://dl.acm.org/citation.cfm?id=1771530.1771554.

9. Sažetak

Cilj ovog rada je dati osnovni uvod u problem klasifikacije videa. Predstavlja se video kao podatak u problemu klasifikacije strojnim učenjem. Predstavljaju se najpoznatiji skupovi za ispitivanje i treniranje. Opisuju se tehnike koje se obično koriste pri izradi modela. Predstavlja se nekoliko poznatih modela koji ostvaruju dobre rezultate na standardnim skupovima za ispitivanje. Predstavlja se način rada povratnih neuronskih mreža i njihove specijalne inačice LSTM-a. Razrađuje se koncept paralelne obrade istih podataka sa modelom "SlowFast Networks". Ukratko se komentiraju rezultati postignuti opisanim modelima i izvodi se zaključak.