Prepoznavanje pisanja s pomočjo konvolucijskih nevronskih mrež in LSTM

Blaž Bertalanič

Fakulteta za elektrotehniko Univerze v Ljubljani, Tržaška 25, 1000 Ljubljana
E-pošta: blaz.bertalanic@protonmail.com

Using CNNs with Long-Short-Term Memory for recognition of writing

Abstract. *This paper addresses the problem of automatic recognition of human behavior, specifically writing, from the image sequences. We propose to use Convolutional Neural Networks utilizing long-short term memory (LSTM) layer to recognize writing of students during a lecture. We present a dataset consisting of more than 13 hours of video involving 22 persons, and describe a structure of the proposed CNN. We present a comparison of recognition accuracy of the classical CNN structure using a single image, and the proposed CNN with LSTM operating on a batch of 10 images. The results indicate that the proposed structure clearly outperforms the baseline method in the detection of writing.*

1 Uvod

Razpoznavanje vzorcev in klasifikacija slik s pomočjo konvolucijskih nevronskih mrež (CNN) [1] v zadnjih letih dosega hiter napredek zahvaljujoč dvigu računskih moči s pomočjo grafičnih kartic. Rešitve, ki uporabljajo to tehnologijo, najdemo praktično na vsakem koraku v našem življenju, od telefonov, do razpoznavanja vzorcev v industriji.

Čeprav konvolucijske nevronske mreže zelo dobro delujejo pri klasifikaciji objektov na slikah (recimo ali je na sliki oseba, pes, mačka, ...), pa pride do težav pri razpoznavanju dogodkov ali akcij, ki so časovno odvisni oziroma trajajo dalj časa. Pri teh je potrebno upoštevati sekvenco zaporednih slik, da lahko podamo končno oceno. To je tudi težava, kateri se bomo posvetili v tem članku.

V članku predstavljamo model dvo-dimenzionalne konvolucijske nevr. mreže v kombinaciji z Long short term memory (LSTM) slojem [2], ki uspešno razpoznava pisanje oseb na video posnetku.

Problem razpoznavanja aktivnosti pisanja je prisoten pri opazovanju in avtomatiziranem ocenjevanju aktivnosti in pozornosti učencev ali študentov v razredu [11], kjer je pisanje ena od pomembnih indikacij. Avtomatizirano razpoznavanje pisanja želimo uporabiti za dolgoročno ocenjevanje aktivnosti in pozornosti učenca v razredu.

2 Pregled literature

Uporaba nevronskih mrež za razpoznavanje aktivnosti ljudi še ni zelo razširjena, saj je potrebno upoštevati časovno komponento.

Skupina indijskih raziskovalcev [3] je uporabila CNN mreže za razpoznavanje gest, vendar so pred slojem LSTM uporabili 3D konvolucijske mreže, ki so precej bolj zahtevne za računanje od naše metode. S svojim modelom so poizkušali razpoznati različne geste ljudi iz video posnetka. Podobna ideja je bila uporabljena tudi za razpoznavo dejanj iz videa zgolj s pomočjo 3D konvolucijskih mrež [4].

Ena od možnosti za analizo video posnetka je tudi kombinacija slik in optičnega toka (ang. Optical flow), kar so uporabili raziskovalci pri razpoznavanju aktivnosti iz video posnetkov [5], želimo pa uporabiti optični pretok v nadaljnjem razvoju prepoznavanja pisanja.

Raziskovalcem je uspelo narediti tudi model, ki je s pomočjo časovne distribucije obdelal vsako sliko videa posebej in na koncu naredil odločitev s polno povezanim slojem [6].

Uporaba 2D konvolucijskih mrež in LSTM je bila uporabljena pri generiranju opisov video posnetkov [7]. Razlika glede na našo metodo je v tem, da pri nas LSTM mreža umesti akcijo v eno izmed dveh skupin, pri njih pa opiše dogajanje na posnetku.

3 Generiranje zbirke podatkov

Zbirka podatkov za prepoznavanje pisanja je bila posneta v laboratoriju LUCAMI na Fakulteti za elektrotehniko. Za namen analize vedenja študentov je bilo izvedeno snemanje študentov s Kinect One senzorjem med spremljanjem lekcije na temo obdelave signalov (DSP) v dolžini 25 minut. V več poskusih je bilo skupno posnetih 22 udeležencev, ki so izpolnjevali teste znanja pred in po lekciji. Analiza vpliva pozornosti na znanje pa je bila tudi objavljena [12].

Video posnetke študentov med lekcijo smo uporabili kot vhodne podatke za učenje modela mreže za prepoznavanje pisanja, pri čemer smo ročno za vsako osebo označili ustrezen razred (pisanje, ne-pisanje) za celotni posnetek. Za prvo skupino šestih oseb smo uporabili celotnih 33 minut video posnetka, medtem ko smo za ostalih 16 uporabili 8 minut dolge izseke iz originalnega video posnetka. Oba video posnetka sta vsebovala 10 sličic na sekundo.

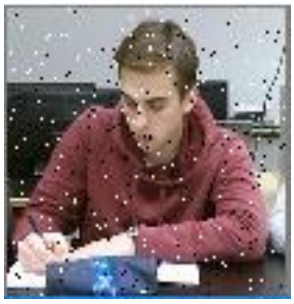
a)

b)

Slika 1. Primer dveh posnetkov predavanj. Prvi dolžine 8 minut in drugi dolžine 33 minut.

Posamezne osebe v video posnetku smo nato izrezali s fiksno pozicijo kvadratnega okna, ki smo ga postavili tako, da je čim bolj zajemal celotno gibanje posamezne osebe. Glede na postavo oseb smo tudi izbrali različne velikosti oken, nato pa pridobljene slike pomanjšali na enotno velikost 100x100 pikslov s tremi RGB kanali. Iz tega smo pripravili dve zbirki podatkov. V prvi smo zaporedne slike združili skupaj v sekvence po 10, kar je v našem primeru pomenilo enosekundni odsek, in dobili končno obliko enega vhodnega podatka (10, 100, 100, 3).

a)



b)

Slika 2. Na sliki a) je primer izreza osebe. Na sliki b) je primer slike z dodanim šumom

Pri drugi zbirki smo slike obdržali samostojno in za referenco poizkušali sistem naučiti prepoznavati pisanje le na slikah (referenčna metoda). Pri obeh zbirkah smo tudi zrcalili vse primere preko vertikalne osi in tako še povečali število primerov. Hkrati smo na tak način poskrbeli tudi, da smo imeli v učnem setu osebe, ki pišejo z levo in desno roko.

Težava pri naših podatkih je bila neenakomerna predstavitev obeh oznak, saj je bilo ne-pisanje prisotno v približno 2/3 vseh podatkov. Ta problem smo rešili na način, da smo primerom pisanja dodali šum in jih nato dodali v zbirko podatkov za učenje. Skupno smo na koncu imeli za približno 13 ur enosekundnih sekvenc pisanja in ne-pisanja, kar je za uporabo v globokem učenju (ang. Deep learning) še vedno relativno majhna zbirka podatkov.

Akcije pisanja smo z ogledi videoposnetkov označili sami za vsako sliko video posnetka, ter nato po 10 zaporednih ocen (na 1 sekundo posnetka) s pomočjo median filtra združili v eno oceno, ki pripada enemu vhodnemu podatku v prvi zbirki in je predstavljala povprečno vrednost pisanja v eni sekundi, medtem ko smo pri drugi zbirki uporabili prvotne oznake.

4 Predlagana metoda

4.1 Oblika mreže

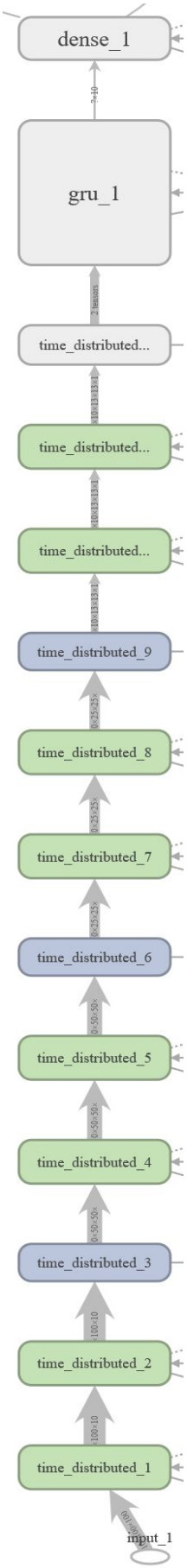
Metode razpoznavanja pisanja smo se lotili na dva načina. Prvi način je napovedoval dejanje na statičnih slikah in smo ga uporabili kot referenco za našo predlagano metodo, medtem ko smo pri drugi metodi upoštevali odvisnost slik med sabo in tako tudi zasnovali mrežo s slojem LSTM, ki nam je pri odločitvi upošteval celotno sekundno sekvenco.

Mreža, ki smo jo uporabili v prvem poskusu (osnovna metoda na posameznih slikah), je imela 13 slojev. Najprej smo postavili tri sklope s po dvema konvolucijskima in enim MaxPooling slojem, temu so sledili Dropout, Flatten in FullyConnected sloj, ki se je nato povezal na izhod, kjer smo dobili našo napoved stanja na sliki.

V drugem primeru (predlagana metoda) je naša mreža ponovno imela tri sklope s po dvema konvolucijskima in enim MaxPooling slojem, ki sta mu sledila še dva konvolucijska sloja, Flatten in nato LSTM sloj, ki pa je bil kasneje zamenjan s slojem Gated recurrent unit (GRU)[8], ki deluje podobno kot LSTM, le da je arhitektura nekoliko preprostejša in posledično olajša računanje, hkrati pa bolje deluje na manjših zbirkah podatkov [9]. Razlog za izbor takšne arhitekture mreže se skriva ravno v zadnjem sloju, saj si LSTM/GRU sloj zapomni prejšne vhodne podatke in jih posledično upošteva pri končni napovedi dejanja. V našem primeru si zapomni vseh 10 slik, ki sestavljajo eno sekundo in tako lahko bolj natančno napove, ali oseba v tem času piše ali ne. Vsi sloji so delovali s časovno distribucijo, saj smo na ta način sekvenco pripeljali do sloja GRU, ki si je nato zapomnil vseh 10 slik in podal končno napoved.

Z uporabo druge metode zmanjšamo možnost, da bi pretentali sistem, ki uporablja statične slike za napovedovanje. Za primer bi lahko oseba držala okvirno pozo na kateri bi delovala kot da piše, v resnici pa roke sploh ne bi premaknila.

Vsi naši konvolucijski sloji so imeli 16 filtrov in vsi, razen v prvem sloju, so bili velikosti 3x3, medtem ko je že omenjeni sloj imel velikost filtrov 7x7. Naš GRU sloj je imel 10 celic. Mreža je bila majhna zaradi majhnega števila oznak in smo se s tem izognili preprileganju in tako dosegli boljšo generalizacijo.



Slika 3. Oblika predlagane mreže za učenje

4.2 Učenje mreže

Mrežo smo učili na Nvidia grafični kartici Titan Xp, ki vsebuje 3840 CUDA jeder s frekvenco delovanja 1.6 GHz in 12 GB grafičnega pomnilnika.

Učenje je potekalo na množici za učenje (ang. Batch) v velikosti 24 vzorcev in z učnim korakom (ang. Learning rate) velikosti 0.000025. Proces je tekel 5 ciklov (ang. Epoch).

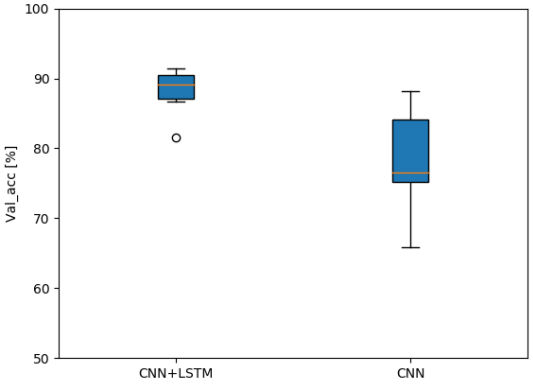
Za optimizacijo smo uporabili SGD algoritem z momentumom 0.9, ki daje najboljše rezultate, če želimo, da je model čim bolj generalen [10].

5 Rezultati

Našo rešitev smo testirali tako, da smo za vse poizkuse izločili vse slike neke osebe iz učne množice podatkov. Oseba, ki smo jo izbrali, je imela označenih 2944 primerov nepisanja in 1036 pisanja za testiranje LSTM metode. Pri navadnih slikah smo imeli 29246 oznak nepisanja in 10554 pisanja.

Slika 4. Primer slike iz validacijskega seta

Ker gre v osnovi še vedno za majhno zbriko podatkov, je bilo naše učenje nekoliko odvisno tudi od sreče, saj se ob začetku procesa uteži določijo naključno in kot take, zaradi manjšega števila podatkov, naš algoritem lahko usmerijo v različne lokalne minimume. Kljub temu se končni rezultati v večini primerov ob ponovitvah razlikujejo le za nekaj odstotkov in imajo povprečno vrednost pravilne napovedi na validacijskem setu 89.1%.



Slika 5. Statistična predstavitev natančnosti (validacijski podatki) pri predlagani metodi CNN+LSTM (levo) in referenčni mreži (desno)

Podobno je bilo z učenjem na navadnih slikah, kjer je bil povprečni odstotek nižji, vendar se nam težava pojavi pri ponavljivosti. V primeru učenja zgolj na posameznih slikah smo zgolj v enem primeru imeli to srečo, da se je model solidno naučil razlikovati med obema akcijama, medtem ko je v ostalih poizkusih večinoma dobro napovedoval ne-pisanje, medtem ko je bilo razmerje med pravilno in nepravilno napovedanimi primeri pisanja okoli polovične uspešnosti.

Za predstavitev rezultatov smo uporabili modela, ki sta na koncu dala najboljši delež pravilnih napovedi na validacijskem setu. Zaradi neenakomerne porazdelitve primerov za pisanje in ne-pisanje pri validacaijskem setu, smo se osredotočili predvsem na uspešnost napovedovanja pisanja.

Izkaže se, da naša metoda bolje napoveduje dejanje pisanja, medtem ko se pri ne-pisanju uspešnost napovedi razlikuje za približno 0.5% in je kot taka praktično nespremenjena.

Tabela 1. Matrika razvrstitev na slikah (primerjalna metoda)

/	Ne-pisanje	Pisanje
Ne-pisanje	**95.6%**	4.4%
Pisanje	32.3%	**67.7%**

Tabela 2. Matrika razvrstitev na sekvenci 10 slik (predlagana metoda)

/	Ne-pisanje	Pisanje
Ne-pisanje	**96.1%**	3.9%
Pisanje	24.2%	**75.8%**

Kot manjši test smo rezultate primerjalne metode združili v enako dolge sekvence, kot smo jih imeli pri testiranju naše metode, in z uporabo median filtra določili skupno napoved na tej sekvenci. Izkaže se, da je rezultat še nekoliko slabši kot na primerjalni metodi.

Naša metoda je izboljšala napoved pisanja kar gre pripisati predvsem temu, da smo z njo upoštevali tudi gibanje oseb, predvsem rok, ki je ključna značilnost tega dejanja. Še vedno pa naša uspešnost ni na dovolj visoki ravni, da bi jo lahko uporabili v splošne namene. Razlog za to gre iskati predvsem v različnih načinih gibanja oseb, drugačnih pozicijah telesa in tudi v različnih slogih pisanja.

6 Zaključek

Pokazali smo, da predlagan model bolje rešuje naš problem kot zgolj navadna CNN mreža. Naša rešitev tako izrablja prednost CNN, ki je razpoznavanje objektov na sliki in hkrati s pomočjo LSTM upošteva časovno odvisnost slik med seboj.

V dosedanjem delu smo se osredotočali zgolj na razpoznavo na video posnetku, zato imamo v nadaljevanju raziskave namen uporabiti tudi globinske slike in izračunan optični tok (ang. Optical flow). Dodatni množici podatkov bi lahko za mrežo predstavljali nove informacije, na podlagih katerih bi potekala končna odločitev ali oseba piše ali ne. Načrtujemo tudi dodatna snemanja, s katerimi bi še povečali velikost in raznolikost učne množice.

Literatura

[1] Y. LeCun and Y. Bengio. »Convolutional networks for images, speech, and time-series,« In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

[2] Sepp Hochreiter and J¨urgen Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, 1997.

[3] Koustav Mullick and Anoop M. Namboodiri, »Learning deep and compact models for gesture recognition,« https://cdn.iiit.ac.in/cdn/cvit.iiit.ac.in/images/ConferencePapers/2017/GESTURE_RECOGNITION.pdf, [Na spletu; dostopano 10. 05. 2018].

[4] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen, “Sign Language Recognition Using Convolutional Neural Networks,” v *European Conference on Computer Vision Workshops*, 2015.

[5] Karen Simonyan and Andrew Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” v *Neural Information Processing Systems*, 2014.

[6] G¨ul Varol, Ivan Laptev, and Cordelia Schmid, “Long-term Temporal Convolutions for Action Recognition,” *arXiv:1604.04494*, 2016.

[7] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell, »Long-term Recurrent Convolutional Networks for Visual Recognition and Description«, arXiv: 1411.4389v4, 2016

[8] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. »On the properties of neural machine translation: Encoder-decoder approaches,« arXiv: 1409.1259, 2014.

[9] Junyoung Chun, Caglar Gulcehr, KyungHyun Cho, Yoshua Bengio, »Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,« arXiv: 1412.3555, 2014.

[10] Nitish Shirish Keskar, Richard Soche, »Improving Generalization Performance by Switching from Adam to SGD,« arXiv: 1712.07628, 2017.

[11] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features". *EURASIP Journal on Image and Video Processing* , ISSN 1687-5176, 2017, vol. 2017, 80, str. 1-12

[12] BURNIK, Urban, ZALETELJ, Janez, KOŠIR, Andrej. Video-based learners' observed attention estimates for lecture learning gain evaluation. *Multimedia tools and applications*, ISSN 1380-7501, 2017.

Zahvala

Rad bi se zahvalil dr. Janezu Zaletelju za vso pomoč pri analizi in reševanju opisanega problema.