Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 621.39
DOI: https://doi.org/10.24867/14BE23Zivkovic

ISTRAŽIVANJE MOGUĆNOSTI SEMANTIČKE SEGMENTACIJE SLIKA SA IOT UREĐAJA

RESEARCH OF SEMANTIC SEGMENTATION CAPABILITIES ON IMAGES FROM IOT DEVICES

Miloš Živković, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIČKO I RAČUNARSKO INŽENJERSTVO

Kratak sadržaj –*Analizirane su mogućnosti semantičke segmentacije slika dobijenih korišćenjem konstruisanog IoT sistema za akviziciju slike, zasnovanog na ESP32-Cam modulu. Implementirane su ERFNet i Unet arhitekture konvolucionih neuronskih mreža. Modeli za segmentaciju obučavani su na Cityscapes i Camvid bazama slika i poređene su njihove performanse. Za testiranje korišćene su slike sa IoT uređaja, DSLR fotoaparata i test slike iz skupova slika korišćenih za obuku modela.*

Ključne reči: *Semantička segmentacija, ERFNet, Unet, IoT uređaji, ESP-32 Cam*

Abstract – *Capabilities of semantic segmentation were analyzed on the images obtained using designed IoT system for image acquisition, based on ESP32-Cam module. ERFNet and Unet convolutional neural network architectures were implemented. Segmentation models were trained on Cityscapes and Camvid image datasets and their performances were compared. For testing were used images from IoT device, DSLR camera, and test images from image datasets utilized for model training.*

Keywords: *Semantic segmentation, ERFNet, Unet, IoT devices, ESP-32 Cam*

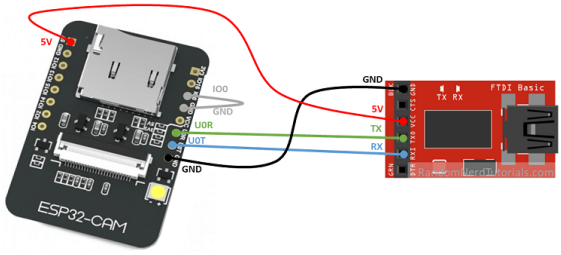
1. UVOD

Klasičan problem kompjuterske vizije je semantička segmentacija čiji je zadatak određivanje klase svakog piksela u slici što vodi ka potpunom razumevanju slike. Najbolje rezultate danas pokazuju metode dubokog učenja, a među njima konvolucione neuronske mreže. Konvolucione neuronske mreže predstavljaju klasu veštačkih neuronskih mreža koje su dizajnirane da automatski i adaptivno nauče prostorne hijerarhije obeležja putem propagacije unazad koristeći nekoliko ključnih blokova kao što su konvolucioni slojevi, udruženi slojevi i potpuno povezani slojevi. Mreže se obučavaju korišćenjem označenih (labeliranih) baza slika. Cilj ovog rada je bilo istraživanje mogućnosti korišćenja IoT (*eng. Internet of Things*) sistema za akviziciju slike kao izvora podataka za primenu algoritama semantičke segmentacije slike. Implementirane su dve arhitekture konv. neuronskih mreža i upoređene njihove performanse.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio doc. dr Branko Brkljač.

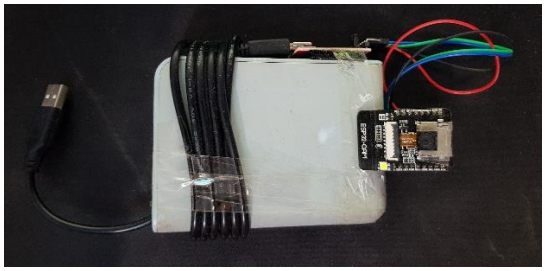
2. ESP32-CAM SISTEM

ESP32-CAM je jeftina ploča za razvoj koja na sebi ima kameru i WiFi modul. Omogućava prenošenje video informacija putem IP adrese u različitim rezolucijama. Za programiranje modula potrebno je koristiti *FTDI* adapter kako bi se omogućila USB konekcija, slika 1. Programira se preko *Arduino* okruženja, a za potrebe ovog rada modul je isprogramiran tako da umesto slanja slike putem IP adrese, sliku snima na *micro SD* memorijsku karticu svaki put kad se pritisne odgovarajuće dugme.



Slika 1. *Povezivanje ESP32-CAM modula i FTDI adaptera*

Kako se modul napaja sa 5V ili 3.3V, nakon programiranja on se može pustiti u rad povezivanjem na računar, eksternu bateriju ili punjač za telefon. Izgled sistema je prikazan na slici 2.



Slika 2. *Izgled ESP32-Cam sistema zajedno sa eksternom baterijom*

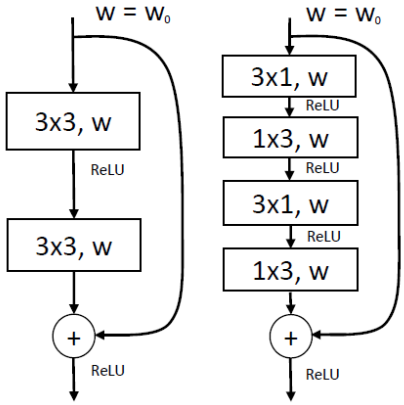
Korišćenjem ovog sistema napravljena je baza slika koja je zajedno sa *Cityscapes, Camvid* i DSLR bazom slika korišćena za proveru performansi treniranih modela.

3. MODELI

Za potrebe rada implementirani su ERFNet i Unet modeli koji su se pokazali kao najefikasniji za rešavanje sličnih problema.

3.1 ERFNet

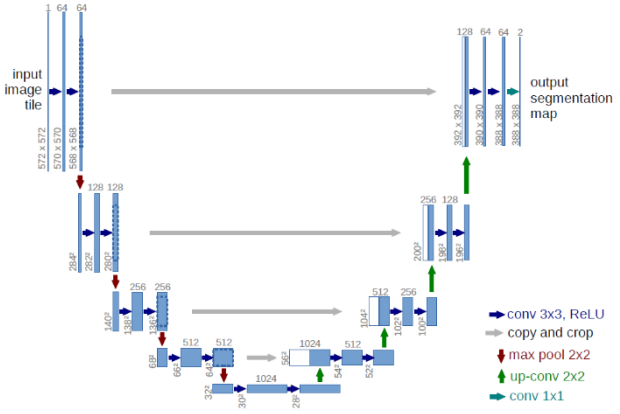
ERFNet [1] je kreiran sa ciljem otklanjanja ograničenja koja ima ResNet [2] arhitektura u svojim ključnim slojevima. Umesto direktnog povezivanja i preskakanja određenih slojeva, ERFNet predlaže redizajniranje sloja bez suženja (*eng. Non-bottleneck*) tako da se u potpunosti koriste konvolucije sa 1D filterima, slika 3.



Slika 3. *Prikaz originalne strukture sloja bez suženja (a) i redizajnirane strukture (b). Preuzeto iz [1]*

3.2 Unet

Arhitektura Unet [3] mreže sastoji se od leve sužavajuće i desne proširujuće strane. Sužavajuća strana prati standardnu arhitekturu konvolucionih mreža i sastoji se od dve 3x3 konvolucije sa ReLU aktivacionom funkcijom i 2x2 slojem maksimalnog udruživanja (*eng. max pooling*). Na svakom koraku smanjenja dimenzionalnosti (*eng. downsampling*) se duplira broj kanala obeležja.



Slika 4. *Arhitektura Unet mreže. Svaki plavi pravougaonik predstavlja višekanalnu mapu obeležja. Broj kanala se nalazi iznad pravougaonika, dimenzije x-y se nalaze kod donjeg levog ugla pravougaonika, dok beli pravougaonici predstavljaju kopirane mape obeležja a strelice predstavljaju različite operacije. Preuzeto iz [3]*

Svaki korak proširujuće strane se sastoji od povećanja dimenzionalnosti (*eng. upsampling*) mapa obeležja (*eng. feature maps*). Nakon čega sledi 2x2 konvolucija koja prepolovi broj kanala obeležja spajanjem sa odgovarajućim odsečenim delom mape obeležja sužavajuće strane mreže, dok se na kraju nalaze dve 3x3 konvolucije sa ReLU aktivacionim funkcijama. Isecanje mapa obeležja je neophodno zbog gubitka graničnih piksela u svakoj konvoluciji. U poslednjem sloju 1x1 konvolucija se koristi za mapiranje svake 64 komponente vektora obeležja u željeni broj klasa. Ukupno, mreža ima 23 konvoluciona sloja, slika 4.

4. BAZE SLIKA

U ovom radu korišćeno je više javno dostupnih baza slika radi treniranja i upoređivanja performansi različitih modela kao i baze slika dobijenih korišćenjem ESP32-Cam sistema i DSLR foto aparata.

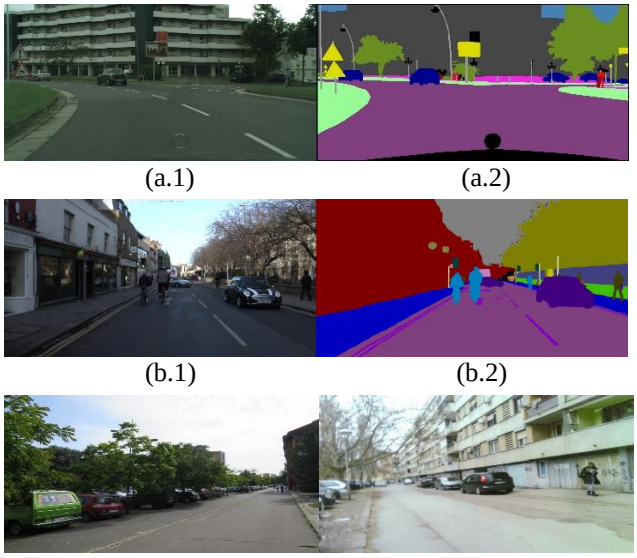
4.1 Cityscapes i CamVid

Cityscapes baza slika [4] predstavlja set slika izvučenih iz video sekvenci snimljenih na ulicama preko 50 različitih gradova. Deo baze korišćen u ovom radu se sastoji od 5000 slika, kao i oznake 35 klasa.

CamVid [5] baza slika takođe se sastoji od seta slika izvučenih iz video sekvence. Baza sadrži 701 par podataka u formi slika - oznaka klase kojoj pripada. Predstavljene su 32 različite klase.

4.2 ESP32-Cam i DSLR

Testiranje pristupa je izvršeno nad bazama napravljenim pomoću DSLR foto aparata i ESP32-Cam sistema. Ove baze ne sadrže oznake klasa i korišćene su radi provere performansi modela treniranih na *Cityscapes* i *CamVid* bazama. Primeri su dati na slici 5.



Slika 5. *Primer slike i slike oznake klasa Cityscapes baze (a.1), (a.2), primer slike i slike oznake klasa CamVid baze (b.1), (b.2), primer slike DSLR baze (c), primer slike ESP32-Cam baze (d)*

5. TRENING I EVALUACIJA MODELA

Biblioteka korišćena za treniranje i evaluaciju modela je preuzeta sa *GitHub* repozitorijuma *tf-semantic-segmentation* [6]. Biblioteka je napravljena specijalno za potrebe semantičke segmentacije i sastoji se od velikog broja alata. Unutar biblioteke postoje definisane arhitekture za najkorišćenije mreže za semantičku segmentaciju, moduli za preuzimanje velikog broja baza slika sa interneta, veliki broj funkcija cene, aktivacionih funkcija, metrika, normalizacionih metoda i metoda za

„augmentaciju“ (proširivanje) skupa podataka. Za potrebe ovog rada kao funkcija cene korišćena je kategorička međuentropija, a kao metrike su korišćeni F1 rezultat, IoU rezultat i prosečna tačnost po klasama.

F rezultat (*eng. F-score*) takođe se naziva F1 [7] rezultat i predstavlja meru tačnosti modela na datom skupu podataka. Definisan je kao harmonijska srednja vrednost preciznosti i odziva modela.

IoU (*eng. Intersection-Over-Union*) [8], takođe poznat kao *Jaccard* indeks, jedan je od najkorišćenijih u oblasti semantičke segmentacije jer je jednostavan a veoma efektivan. IoU rezultat predstavlja preklapanje rezultata modela sa označenom slikom podeljen unijom između rezultata modela i označene slike. IoU rezultat ima raspon vrednosti od 0 do 1 i njegova vrednost tokom obuke prikazana je na slici 7.

ERFNet i Unet modeli trenirani su posebno na *Cityscapes* i na *CamVid* bazama slika i u svim slučajevima podaci su distribuirani u odnosu 80%, 10% i 10% za trening, validaciju i test, respektivno. Kao funkcija cene je korišćena kategorička međuentropija (*eng. categorical crossentropy*), aktivaciona funkcija je bila softmaks (*eng. softmax*), i korišćen je Adam optimajzer (*eng.Adam optimizer*).

Treniranje svakog od modela rađeno je od samog početka, odnosno bez korišćenja pretreniranog modela. Broj epoha za treniranje i vreme za koje svaki od modela generiše izlaz za ulaznu sliku su dati u tabeli 1.

Model (trening skup)	Broj epoha	Vreme izvršavanja (s)
ERFNet (Cityscapes)	47	11,3
ERFNet (CamVid)	98	9,7
Unet (Cityscapes)	38	310,8
Unet (CamVid)	52	284,6

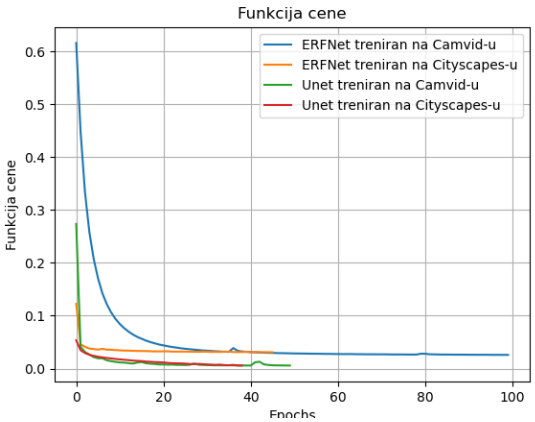
Tabela 1. *Broj epoha za treniranje i vreme izvršavanja treniranog modela*

Kategorička međuentropija je funkcija cene koja se koristi za višeklasne probleme, odnosno one probleme gde jedan uzorak (piksel) može pripadati samo jednoj od više mogućih kategorija (klasa) i model treba da odluči kojoj od njih pripada [6]. Vrednost funkcije cene je prikazana na slici 6.

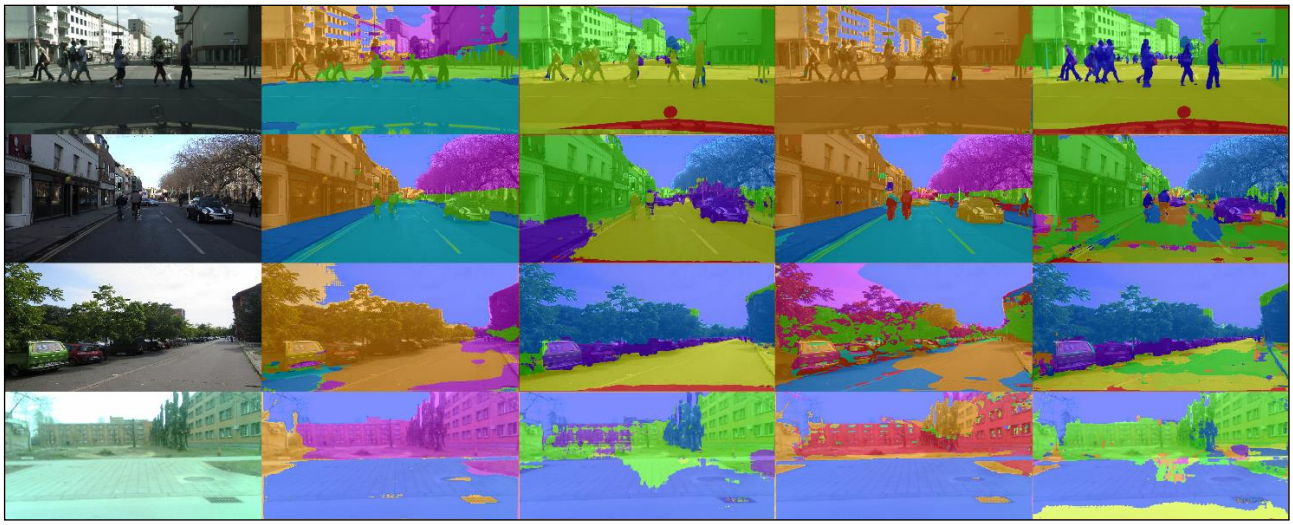
6. REZULTATI

Bolje performanse prilikom treniranja postignute su korišćenjem Unet arhitekture u slučaju obe baze slika. Najlošije se pokazala ERFNet arhitektura trenirana na *Cityscapes* bazi slika.

Sva četiri modela su trenirana različitim brojem epoha. Trening se zaustavio ukoliko se funkcija cene nije poboljšala, odnosno smanjila deset uzastopnih epoha. Najkraće je trenirana Unet arhitektura na *Cityscapes* bazi slika, a najduže ERFNet arhitektura na *CamVid* bazi slika. Prosečna tačnost po klasama na trening skupu je data na slici 8. Vizuelna poređenja rezultata modela semantičke segmentacije za ulazne slike su data na slici 9.



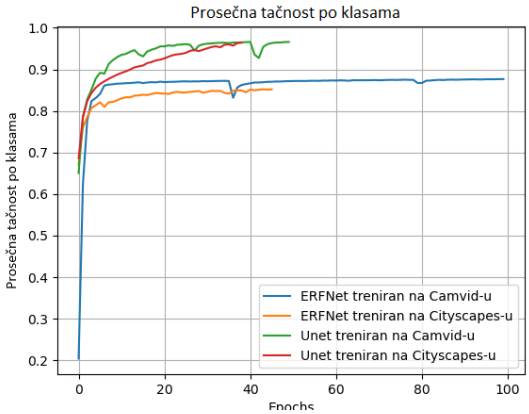
Slika 6. *Vrednosti funkcije cene prilikom treninga modela*



(a) (b) (c) (d) (e)

Slika 9. *Redovi predstavljaju slike iz različitih baza: prvi red odgovara slici iz Cityscapes baze, drugi red slici iz CamVid baze, treći red slici iz DSLR baze i četvrti red slici iz ESP32-Cam baze, dok kolone 2-5 predstavljaju rezultate semantičke segmentacije za: ERFNet treniran na CamVid (b), ERFNet treniran na Cityscapes (c), Unet treniran na CamVid (d) i Unet treniran na Cityscapes (e). Prva kolona (a) odgovara ulaznim slikama.*

Slika 7. *Vrednosti IoU rezultata prilikom treninga modela*



Slika 8. *Vrednosti prosečne tačnosti po klasama prilikom treninga modela*

7. ZAKLJUČAK

U ovom radu opisani su i diskutovani načini za rešavanje problema semantičke segmentacije slike. U istraživanju su korišćeni skupovi podataka koji odgovaraju različitim senzorima slike uključujući i kamere namenjene korišćenju u kombinaciji sa IoT uređajima male potrošnje energije (*ESP32-CAM modul).* Predloženo IoT rešenje za akviziciju slike je implementirano korišćenjem odgovarajućeg hardvera i softvera koji su razvijeni kao deo istraživanja u okviru master rada. Radi demonstracije rada mobilne platforme za akviziciju slike koja ima malu potrošnju energije i kao takva je pogodna za primene na otvorenom prostoru, prikupljena je baza slika urbanog okruženja na različitim lokacijama u Novom Sadu. Na prikupljenoj bazi slika su merene performanse nekoliko modela za semantičku segmentaciju slike (*ERFNet, Unet*). Pomenuti modeli su obučavani (trenirani) na odgovarajućim skupovima podataka (*Camvid, Cityscapes*).

U okviru istraživanja i rada na tezi izvršena je kompletna obuka modela. Rezultati obuke različitih modela prikazani su u poglavlju 5, kao i vizuelni prikazi rezultata segmentacije na nasumično izabranim slikama iz više korišćenih baza uključujući i baze kreirane u okviru ovog rada (*DSLR i ESP32-CAM baze*). Uzimajući u obzir potencijalna ograničenja IoT uređaja, ERFNet je od početka odabran kao model manje računske složenosti koji bi mogao da pruži uporedive rezultate kao i neki složeniji modeli. Tokom obuke ERFNet je pokazao lošije vrednosti ciljne funkcije i ranije je ušao u zasićenje, slika 7. Slično i prilikom merenja performansi na test podskupu, imao je lošiji rezultat u odnosu na predstavnika složenijih modela kao što je *Unet*. Ovo može biti posledica manje složenosti modela, kao i prevelikog prilagođavanja podacima u trening skupu. Ovo je posebno uočljivo u slučaju veće baze slika kao što je *Cityscapes (*u oba razmatrana scenarija – kada je *ERFNet* obučavan na *Camvid*, a testiran na *Cityscapes* i obrnuto) dok na jednostavnijoj bazi slika kao što je *Camvid* (za koju je *ERFNet* originalno predložen i obučavan), vizuelni prikaz rezultata u drugom redu i drugoj koloni slike 9 demonstrira da ERFNet može da postigne dobre rezultate na *Camvid* skupu na kome je i treniran.

Opisano ponašanje bi trebalo dodatno istražiti kako bi se utvrdilo da li je manja mogućnost generalizacije (uopštavanja) rezultat: 1) manje kvalitetnog skupa za obuku (kao kada je *ERFNet* obučavan na *Camvid*, a testiran na *Cityscapes*), ili 2) nedovoljno prilagođenih hiper parametara tokom obuke modela sa drugim bazama slika (npr. kada je *ERFNet* obučavan na *Cityscapes*, a testiran na *Camvid*). Generalni zaključak jeste da modeli obučeni na određenom skupu slika na istom takođe pokazuju bolje rezultate, što je i očekivano.

8. LITERATURA

[1] Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R. (2017). ErfNet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems.

[2] He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep residual learning for image recognition. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

[3] Ronneberger, O. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Int. Conf. on Medical image computing and computer-assisted intervention*.

[4] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding, *CVPR*.

[5] Fauqueur, J., Brostow, G., Cipolla, R. (2007). Assisted video object labeling by joint tracking of regions and keypoints, *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil.

[6] https://github.com/baudcode/tf-semantic-segmentation

[7] Lipton, Z. C., Elkan, C., Narayanaswamy, B. (2014). Optimal thresholding of classifiers to maximize F1 measure, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

[8] Rezatofighi, H., Tsoi, N, Gwak, J., Sadeghian, A., Reid, I., Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression, *CVPR*.

Kratka biografija:

Miloš Živković rođen je u Valjevu 1996. god. Master rad na Fakultetu tehničkih nauka iz oblasti Energetika, elektronika i telekomunikacije – Obrada signala, odbranio je 2021.god. Osnovne akademske studije završio je 2019. godine na studijskom programu Biomedicinsko inženjerstvo.
kontakt: pho.milos@gmail.com