Semantička segmentacija prometnih slika pomoću modela za predviđanje u stvarnom vremenu

Dominik Stipić
prof. dr. sc. Siniša Šegvić

Siječanj 2021

1 Uvod

Većina modela semantičke segmentacije razvijena je s ciljem poboljšanja točnosti raspoznavanja. Takav pristup uzrokovao je to da većina modela ima veliku memorijsku složenost, dugo se uče i predviđanje im traje jako dugo. Semantička segmentacija objekata je jedna od ključnih metoda koje je potrebno razviti prilikom izrade autonomnih vozila. Takva funkcionalnost izvodit će se na ugrađenim sustavima čiji je memorijski kapacitet bitno manji od memorija računala na kojima se modeli uče. Još jedan problem javlja se zbog toga što je vožnja zadatak u kojem je potreban vrlo brzi odziv na promjene u okolini. Zbog tog je razloga brzina ključna i moramo imati modele koji će biti sposobni raditi u stvarnom vremenu. Moramo biti svjesni da ubrzavanjem i smanjenjem memorijske složenosti smanjujemo kompleksnost modela, a samim time i njegove performanse. Iz tog razloga javlja se kompromis između točnosti- brzine i složenosti modela.

U ovom projektu opisane su dvije inačice modela za rad u stvarnom vremenu i napravljene su evaluacije tih inačica na CamVid skupu podataka, koji je standardni skup podataka za semantičku segmentaciju objekata iz prometa.

2 Skup za učenje - CamVid

CamVid je označeni skup slika razvijen na Cambridgeu[1]. Skup podataka se sastoji od prometnih slika, a objekti na slici su označeni u 32 semantičke klase. Slike su dobivene snimanjem okoline tijekom vožnje automobila s ciljem što vjernijeg oslikavanja situacija u kojima se može naći vozač. U konačnici je dobiveno otprilike 700 slika koje su potom ručno označene u 32 razreda, a onda je njihova točnost ponovno testirana. Svi razredi CamVid skupa podataka vidljivi su na slici 1.

U praksi su mnogi modeli naučeni i testirani na samo 11 najčešćih razreda i Void razredu. Dobiveni skup podataka s 12 razreda je podijeljen na 3 podskupa:

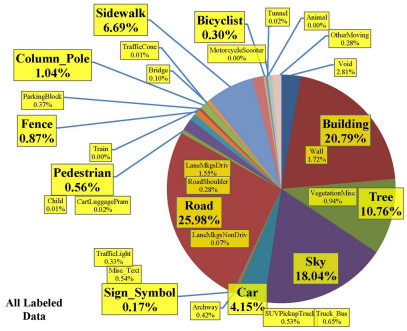

Figure 1: Udio pojedinih oznaka u CamVid skupu. Slika preuzeta sa službene stranice

skup za učenje, skup za validiranje i skup za testiranje. Veličine podskupa su:

- Skup za učenje -> 369 primjera
- Skup za validiranje -> 100 primjera
- Skup za testiranje -> 232 primjera

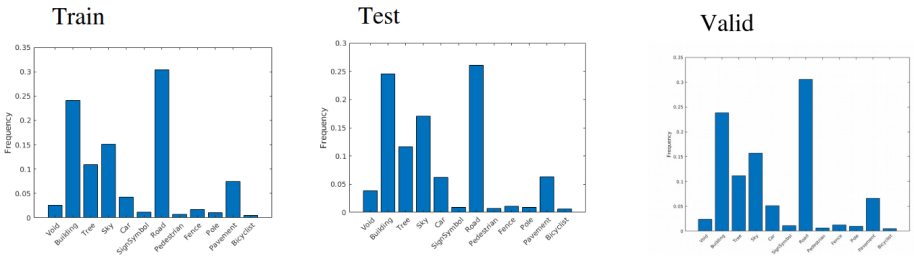

Figure 2: Histogram frekvencija za skup za učenje, testiranje, validiranje

3 Model za semantičku segmentaciju - SwiftNet

SwiftNet je arhitektura namijenjena za semantičku segmentaciju prometnih slika koja je napravljena s ciljem predviđanja razreda objekta u stvarnom vremenu[4]. Dizajn SwiftNeta je vođen s ciljem da se može izvoditi na mobilnim i ugradbenim uređajima ograničenih performansi. SwiftNet je predstavljen u dvije inačice, u prvoj inačici korištena je prostorno sažimajuća piramida (*engl. spatial pyramid pooling - SPP*)[3], a u drugoj inačici korištena je metoda rezolucijskih piramida (*engl. resolution pyramid*). Cilj obje inačice je povećanje receptivnog polja modela što doprinosi boljoj detekciji velikih objekta na slici. Tip arhitektura

modela je koder-dekoder arhitektura s lateralnim vezama između dekodera i enkodera. Enkoder kao izlaz daje sažetu i semantički bogatiju verziju ulazne slike, te zbog toga razloga pikseli manjih objekta imaju nižu vjerojatnost prepoznavanja na izlazu. Rješenje tog problema jest kombiniranje aktivacija plićih slojeva i dubljih slojeva. U SwiftNet implementaciji enkoder je izveden pomoću rezidulanih blokova (*engl. resnet*)[2].

Arhitekture SwiftNeta sa SPP modulom (*Single Scale SwiftNet*) i piramidalnog SwiftNeta prikazane su na slikama 3 i 4.

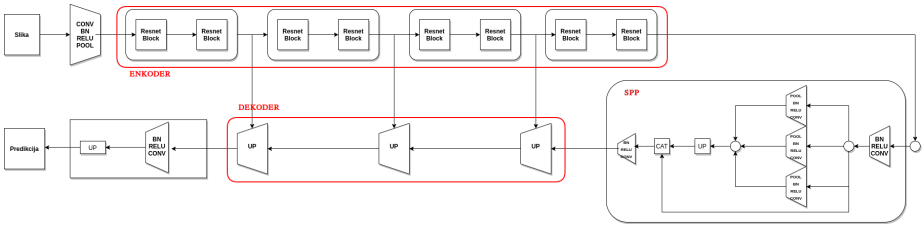

Figure 3: Single Scale SwiftNet, crvenom bojom označeni su moduli enkodera i dekodera. Između enkodera i dekodera nalazi se SPP modul, a na kraju modela nalazi se slojevi za fino podešavanje modela na zadani skup podataka.

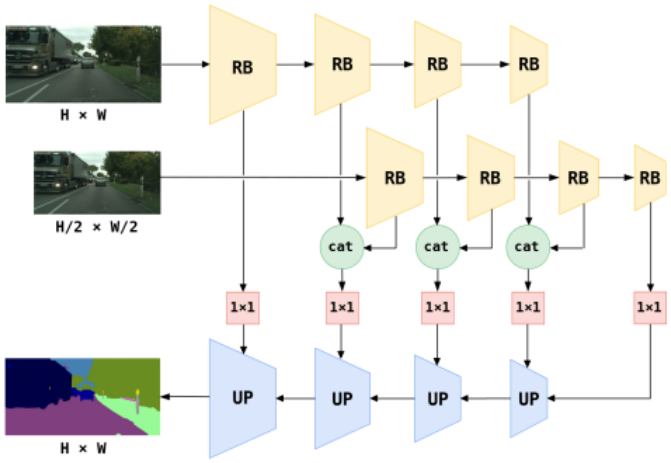

Figure 4: SwiftNet s 2-razinskom piramidom. K-razinska piramida interpolira slike na k različitih rezolucija što rezultira povećanjem receptivnog polja aktivacija. Parametri enkodera su dijeljeni. Značajke istih rezolucija se spajaju lateralnim vezama i kombiniraju sa semantički bogatijim značajkama.

4 Eksperimenti i rezultati

Eksperimenti su napravljeni s obje inačice arhitekture, učenje modela se radilo na Google Colab platformi i na njihovim grafičkim karticama. Modeli su se učili kroz 400 epoha sa stopom učenja $\eta = 0.0004$, veličinom grupe (*engl. batch size*) od 12 primjera i sa snižavanjem težina (*engl. weight decay*) od 0.0001. Za optimiranje parametra korišten je Adam optimizator, a za promjenu stope učenja kroz epohe bio je odgovoran CosineAnnealingLR Pytorch modul s periodom od 250 epoha i minimalnom stopom učenja od 10^{-6}. Funkcija pogreške bila je unakrsna entropija.

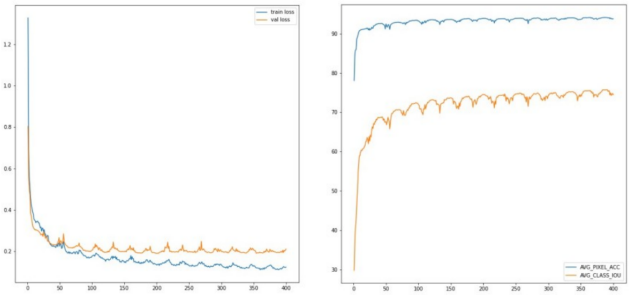

Figure 5: **Jednorazinski SwiftNet**. Na lijevom grafu prikazano je kretanje funkcije pogreške na skupu za učenje i skupu za validaciju kroz epohe, dok je na desnom grafu prikazano kretanje točnosti i MIoU metrika kroz epohe na skupu za validiranje.

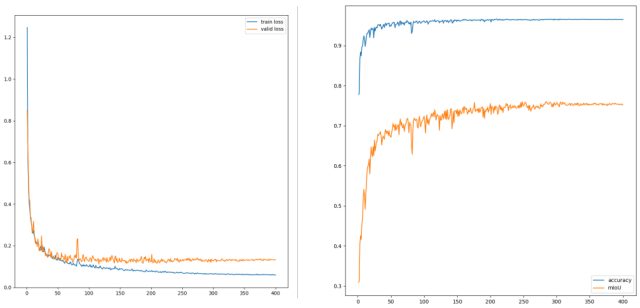

Figure 6: **Piramidalni SwiftNet**. Na lijevom grafu prikazano je kretanje funkcije pogreške na skupu za učenje i skupu za validaciju kroz epohe, dok je na desnom grafu prikazano kretanje točnosti i MIoU metrika kroz epohe na skupu za validiranje.

Model	Točnost	mIoU
Single scale	93,08%	72,80%
Pyramid	92,67%	73,31%

Figure 7: Rezultati nakon evaluacije na skupu za testiranje.

	Single scale	Pyramid
Building	86,00%	85,85%
Tree	78,79%	81,53%
Sky	92,92%	92,71%
Car	84,07%	80,00%
Sign	55,23%	53,94%
Road	96,69%	95,69%
Pedestrian	55,28%	58,69%
Fence	73,91%	72,43%
Column Pole	37,37%	36,45%
Sidewalk	83,55%	80,82%
Bicyclist	57,02%	68,29%

Figure 8: mIoU mjera za svaki razred.

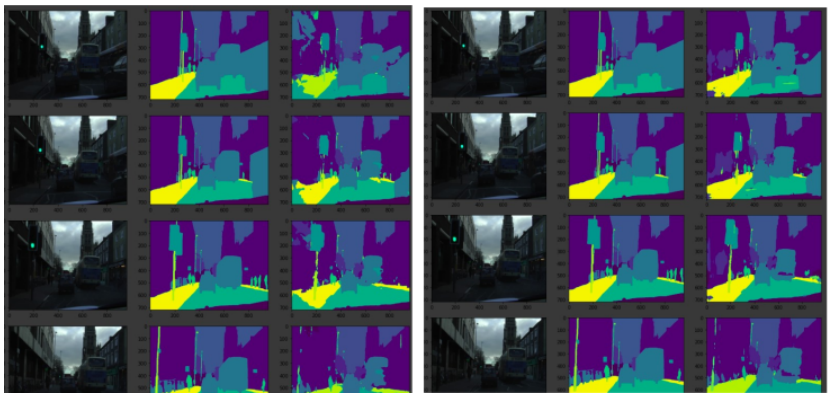

Figure 9: Predikcije piramidalnog SwiftNeta(lijevo) i jednorazinskog SwiftNeta (desno) na 4 slike. U prvom stupcu nalazi se ulazna slika, u drugom stupcu nalaze se točne oznake, a u posljednjem stupcu nalaze se predikcije modela

5 Literatura

References

[1] G. Brostow, J. Fauqueur, and R. Cipolla. "Semantic object classes in video: A high-definition ground truth database". In: *Pattern Recognit. Lett.* 30 (2009), pp. 88–97.

[2] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[3] Kaiming He et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". In: *CoRR* abs/1406.4729 (2014). arXiv: 1406.4729. URL: http://arxiv.org/abs/1406.4729.

[4] Marin Orsic et al. "In Defense of Pre-trained ImageNet Architectures for Real-time Semantic Segmentation of Road-driving Images". In: *CoRR* abs/1903.08469 (2019). arXiv: 1903.08469. URL: http://arxiv.org/abs/1903.08469.