Automatsko prepoznavanje teksta sa tablica automobila koristeći Jedinstveni Vizuelni Model za Prepoznavanje Teksta na Sceni

Student: Andrija Urošević

Mentor: dr Nemanja Ilić

Računarski fakultet,
Univerzitet Union

Jul 2024

Predgovor

Prostor za predgovor.

Sadržaj

Sažetak

Automatsko prepoznavanje teksta sa registarskih tablica automobila od izuzetne je važnositi za savremene sisteme nadzora saobraćaja, praćenje vozila i obezbeđenje sigurnosti na putevima. Identifikacija registarskih tablica ima različite primene, uključujući praćenje ukradenih vozila, naplatu putarine, sigurnosne provere i nadzor saobraćaja. U proteklim decenijama, napredak u oblasti obrade slike i mašinskog učenja omogućio je razvoj efikasnih sistema za automatsko prepoznavanje teksta sa tablica vozila. Primena dubokih neuronskih mreža i algoritama dubokog učenja omogućila je visoku tačnost prepoznavanja teksta, čak i u složenim scenama i različitim uslovima snimanja. Ovaj rad istražuje metode i tehnike za automatsko prepoznavanje teksta sa tablica automobila, sa ciljem razvoja sistema koji može precizno identifikovati registarske tablice u realnom vremenu. Eksperimentalni rezultati prikazuju performanse sistema u stvarnim uslovima i ukazuju na mogućnosti za primenu u različitim oblastima, uključujući nadzor saobraćaja, bezbednosne provere i identifikaciju vozila.

1 Uvod

Automatsko prepoznavanje teksta je ključna tehnologija u oblasti kompjuterske vizije koja ima široku primenu u različitim aplikacijama, uključujući prepoznavanje registarskih tablica vozila, prepoznavanje rukopisa, prepoznavanje dokumenata i mnoge druge. Glavni cilj automatskog prepoznavanja teksta je pretvaranje vizuelno prikazanog teksta u format koji računari mogu razumeti i obrađivati, omogućavajući im da interpretiraju tekstualne informacije slično kao što to radi čovek.

Automatsko prepoznavanje teksta sa registarskih tablica automobila obuhvata nekoliko ključnih koraka koji se odvijaju u procesu od prikupljanja podataka do konačne integracije sistema u softver za prepoznavanje tablica automobila.

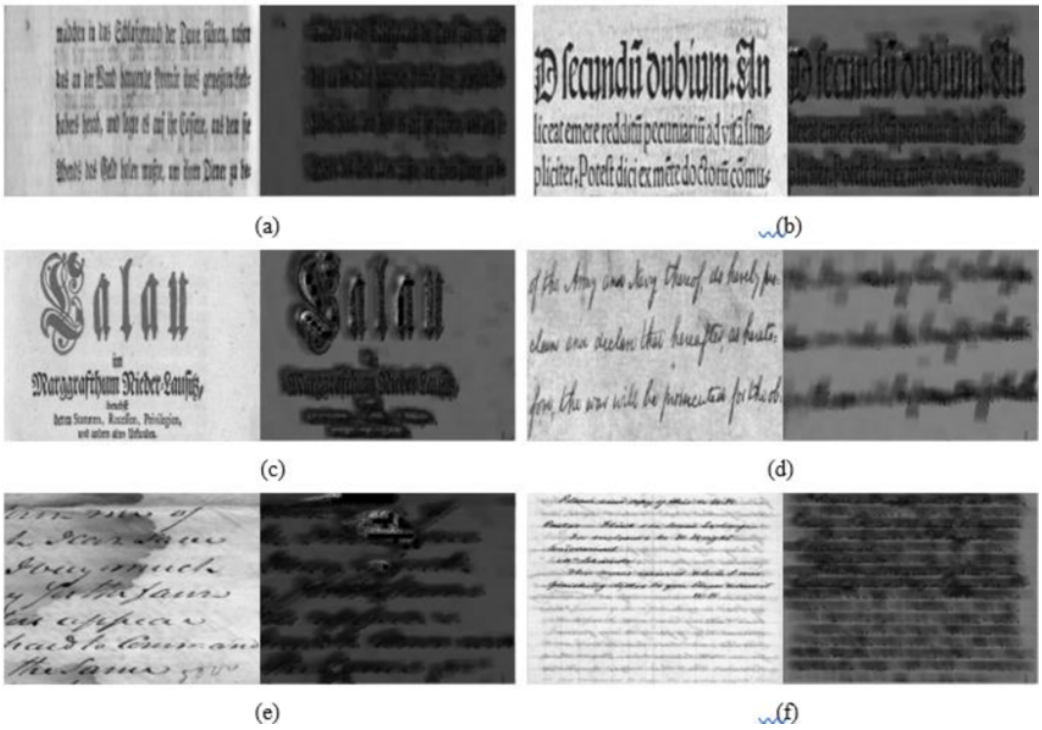
Prikupljanje raznovrsnog skupa slika registarskih tablica vozila ključno je za uspešno treniranje modela. Ove slike treba da obuhvataju različite tipove tablica, različite uslove osvetljenja i pozadine kako bi model bio što robustniji. Nakon prikupljanja, slike treba pažljivo razvrstati na one koje su pogodne za treniranje modela i one koje nisu. Ovo uključuje filtriranje slika sa veoma lošim kvalitetom, zamućenim ili nejasnim tablicama. Kako bi se obogatio skup podataka i poboljšala generalizacija modela, potrebno je primeniti tehnike augmentacije podataka. Ovo uključuje manipulaciju sa slikama kao što su rotacija, promena osvetljenja, izobličenja i dodavanje šuma. Pored toga, sintetički podaci se mogu generisati korišćenjem programa za generisanje tablica sa tekstom. Svaka slika mora biti precizno označena sa tačnim tekstualnim sadržajem registarske tablice kako bi se koristila za obuku modela. Ovaj proces može biti ručan ili se može koristiti alat za automatsko lejbelovanje. Nakon pripreme podataka, sledi faza treniranja mreže za prepoznavanje teksta. U ovoj fazi, model se obučava nad označenim podacima kako bi naučio da prepoznaje tekst sa slika tablica. Kada je model obučen, integriše se u softver za prepoznavanje tablica automobila. Ovaj softver obično obuhvata module za detekciju tablica, detekciju teksta na tablicama, formatiranje izlaza i druge funkcionalnosti.

Kako bi omogućili portabilnost i lakšu distribuciju sistema za prepoznavanje tablica, koristi se Docker kontejner. Docker omogućava pakovanje softverskih aplikacija i njihovo pokretanje u izolovanim okruženjima. Još jedna od bitnih komponenata je Python web framework - FastAPI koji omogućava brzo kreiranje API-ja za komunikaciju sa softverskim komponentama. Integracija Docker-a i FastAPI modula omogućava da se servis za prepoznavanje teksta koristi nezavisno od platforme na kojoj se izvršava, čineći ga pristupačnim i jednostavnim za upotrebu u različitim okruženjima.

2 Prepoznavanje teksta

2.1 Uvod u prepoznavanje teksta

Prepoznavanje teksta na sceni ima za cilj da tekst sa slika konvertuje u digitalni niz karaktera, što prenosi semantiku visokog nivoa ključnu za razumevanje scene. Zadatak prepoznavanja teksta sa scene je izazovan zbog varijacija u: deformacijama teksta, fontovima, preklapanjima različitih tekstova, prekompleksnim pozadinama, itd. Dodatno, otežavajući faktor može biti i to što se tekst može pojaviti pod različitim uglovima.



Slika 1: Tekst na sceni sa različitim fontovima, pozadinama, osvetljenjem i slično

U proteklim godinama uloženi su brojni napori kako bi se poboljšala tačnost prepoznavanja teksta. Moderni pristupi za prepoznavanje teksta, pored tačnosti, takođe uzimaju u obzir i faktore poput brzine izvršavanja modela zbog praktičnih zahteva.

Metodološki, prepoznavanje teksta na sceni može se posmatrati kao prelazak iz modaliteta slike u niz karaktera. Obično, prepoznavanje teksta se sastoji od dva osnovna dela, vizuelnog modela za ekstrakciju karakteristika i sekvencijskog modela za transkripciju teksta.

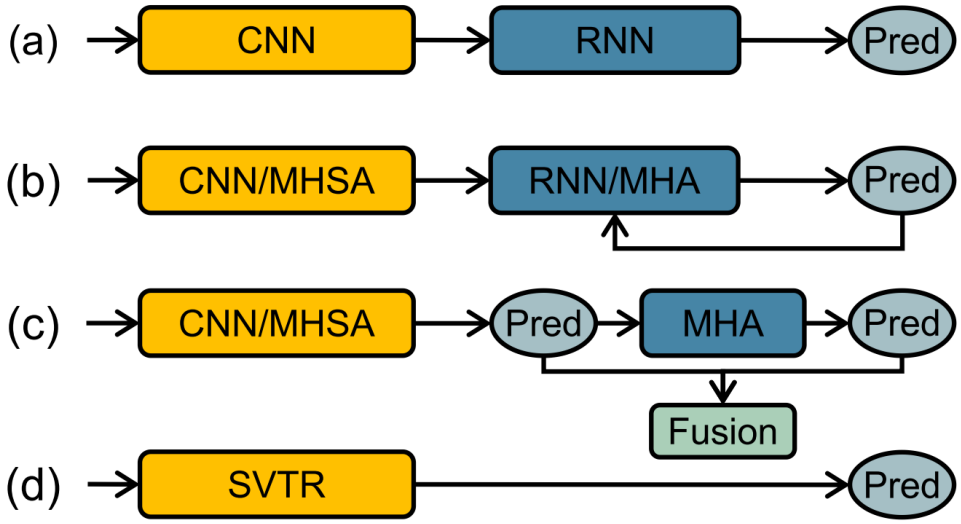
2.2 Istorijski pregled prepoznavanja teksta

Prvi primeri i prva faza tehnologije optičkog prepoznavanja karaktera(OCR) pojavili su se sredinom 20. veka, pretežno tokom 1950-ih i 1960-ih godina. Ovo doba obeležilo je razvoj ranih sistema OCR-a, koji su koristili osnovne tehnike prepoznavanja obrazaca kako bi prepoznali mašinski odštampane karaktere. Ovi sistemi su često bili ograničeni na prepoznavanje određenih fontova i imali su relativno nisku stopu tačnosti u poređenju sa modernom OCR tehnologijom. Glavna primena im je bila čitanje standardizovanih obrazaca i dokumenata sa jasno štampanim tekstom i poznatim fontom.

Druga faza tehnologije optičkog prepoznavanja karaktera dogodila se krajem 20. veka i početkom 21. veka, počevši oko 1970-ih i nastavljajući se u 2000-ima. Ovo doba je obeleženo značajnim napretkom u tehnologiji OCR-a, uključujući razvoj sofisticiranih algoritama i tehnika za prepoznavanje karaktera. Ovi napredci doveli su do veće tačnosti i mogućnosti prepoznavanja šireg spektra fontova, jezika i rasporeda dokumenata. Dodatno, integracija pristupa mašinskog učenja i neuronskih mreža doprinela je daljem poboljšanju performansi OCR-a. U ovoj fazi došlo je do primene OCR sistema u širem spektru aplikacija, od skeniranja i konverzije dokumenata u digitalno arhiviranje, do automatizovanog unosa podataka i ekstrakcije teksta u različitim industrijama.

Treća faza tehnologije optičkog prepoznavanja karaktera je trenutno aktuelna i predstavlja trenutno stanje napretka u sistemima OCR-a. Ovo doba karakteriše integracija najnovijih tehnologija poput dubokog učenja, konvolucionih neuronskih mreža (CNN) i rekurentnih neuronskih mreža (RNN) u algoritme OCR-a. Ove napredne tehnike značajno su poboljšale tačnost i pouzdanost sistema OCR-a, omogućavajući prepoznavanje složenih dokumenata sa različitim fontovima, rasporedima i jezicima. Osim toga, pojava OCR usluga zasnovanih na cloud-u i integracija OCR funkcionalnosti u mobilne uređaje učinili su OCR dostupnijim i svestranijim nego ikad ranije. Treća faza takođe obuhvata napretke u analizi i razumevanju dokumenata, omogućavajući OCR sistemima da izvlače ne samo tekst već i strukturalne i semantičke informacije iz dokumenata, što dovodi do poboljšanih sposobnosti obrade dokumenata i pretraživanja informacija.

2.3 Arhitekture modela za prpoznavanje teksta na slici



Slika 2: Arhitekture modela za prepoznavanje teksta sa scene. (a) Modeli zasnovani na CNN-RNN. (b) Modeli kodiranja-dekodiranja. (c) Vizuelno-jezički modeli. (d) SVTR, koji prepoznaje tekst scene sa jedinstvenim vizuelnim modelom i odlikuje se efikasnošću, tačnošću i višejezičnom svestranošću.

Modeli zasnovani na CNN-RNN [SBY15] prvo koriste CNN za ekstrakciju karakteristika. Karakteristike se zatim preoblikuju u sekvencu koju BiLSTM modeluje uz pomoć CTC gubitka kako bi generisao predikciju (Slika 2(a)). Odlikuju se efikasnošću i ostaju izbor za neke komercijalne proizvode za prepoznavanje teksta sa scene. Međutim, preoblikovanje karakteristika u sekvencu je osetljivo na deformacije teksta, što ograničava efikasnost takvih modela.

Kasnije su pristupi zasnovani na auto-regresivnim metodama kodera-dekodera postale popularne [SCX19; Li+19; Zhe+23]. Ove metode transformišu prepoznavanje u iterativni proces dekodiranja (Slika 2(b)). Kao rezultat, postignuta je poboljšana tačnost jer je uzeta u obzir kontekstualna informacija. Međutim, brzina izvođenja je spora zbog transkripcije karakter po karakter. Ovaj postupak je dodatno proširen na softversku strukturu zasnovanu na viziji i jeziku [Yu+20; Fan+21], gde je jezičko znanje uključeno (Slika 2(c)) i sprovedena je paralelna predikcija. Ipak, ovaj postupak često zahteva model sa velikim kapacitetom ili složenu paradigmu prepoznavanja kako bi se osigurala tačnost prepoznavanja, što ograničava njegovu efikasnost.

U poslednje vreme, naglasak je na razvoju pojednostavljenih arhitektura kako bi se dobilo na brzini izvršavanja. Na primer, korišćenje složene paradigme obuke, ali jednostavnog modela za izvršavanje. Rešenje zasnovano na CRNN-RNN ponovo je pregledano u sledećem radu: [Hu+20]. Koristi mehanizam pažnje i grafovsku neuronsku mrežu za agregaciju sekvencijalnih karakteristika koje odgovaraju istom karakteru. Pri izvršavanju, deo za modelovanje zasnovan na mehanizmu pažnje je odbačen kako bi se uskladila tačnost i brzina.

Nedavni uspeh transformera za obradu slike [Dos+21; Liu+21], inspirisao je nastanak jedinstvenog vizuelnog modela za prepoznavanje teskta na sceni(SVTR) [Du+22]. SVTR najpre razlaže tekst slike na male 2D isečke koji se nazvaju komponente karaktera, od kojih svaka komponenta može sadržati samo deo karaktera. Tokenizacija slike po isečcima praćena mehanizmom samopažnje se primenjuje da bi se uhvatile indicije prepoznavanja teksta među komponentama karaktera. Za ovu svrhu je razvijena prilagođena arhitektura za tekst, čija osnovna struktura sadrži progresivno smanjujuću visinu mape karakteristika u tri faze i uključujue operacije mešanja, spajanja i/ili kombinovanja. Osmišljeni su lokalni i globalni blokovi mešanja koji se rekurzivno primenjuju u svakoj fazi, zajedno sa operacijom spajanja ili kombinovanja, stičući tako afinitete na nivou lokalnih komponenti koje predstavljaju karakteristike slične potezima karaktera i dugoročne zavisnosti između različitih karaktera. Dakle, osnovna struktura ekstraktuje karakteristike komponenti na različitim rastojanjima i na više skala, formirajući višeslojnu percepciju karakteristika karaktera. Kao rezultat, prepoznavanje teksta se postiže jednostavnom linearnom predikcijom. U celom procesu koristi se samo jedan vizuelni model (Slika 2(d)).

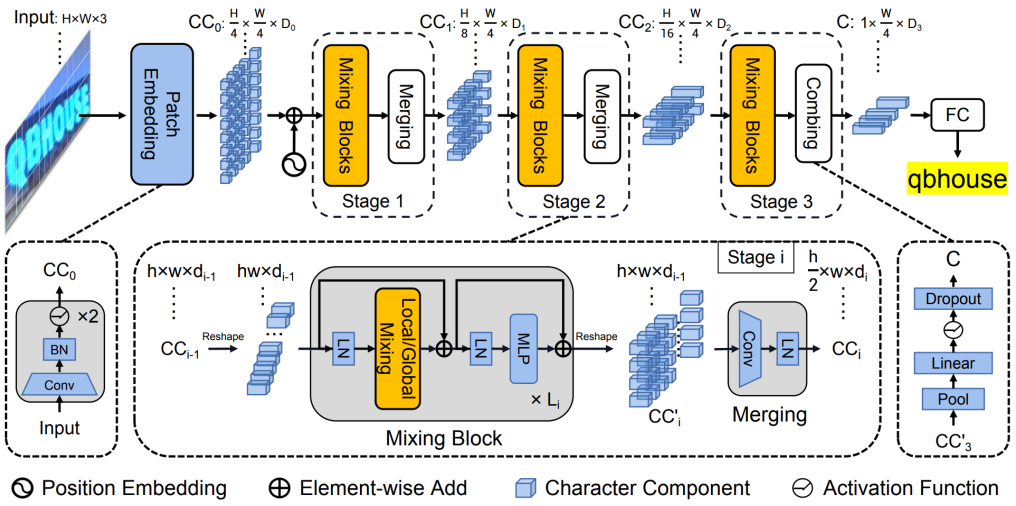
2.4 Korišćenje Jedinstvenog Vizuelnog Modela za Prepoznavanje Teksta na Sceni

Tradicionalni modeli za prepoznavanje teksta obično uključuju dve odvojene komponente: vizuelni model za izdvajanje karakteristika sa slike i sekvencijalni model za dekodiranje izdvojenih karakteristika u tekst. Jedinstveni vizuelni model za prepoznavanje teksta na sceni(SVTR) eliminiše potrebu za sekvencijalnim modelom u potpunosti, čineći ga jednostavnijim i efikasnijim.

Uklanjanjem komponente sekvencijalnog modeliranja, SVTR postiže konkurentnu preciznost na zadacima prepoznavanja teksta, pružajući veću efikasnost u poređenju sa tradicionalnim metodama.

2.4.1 Arhiterktura

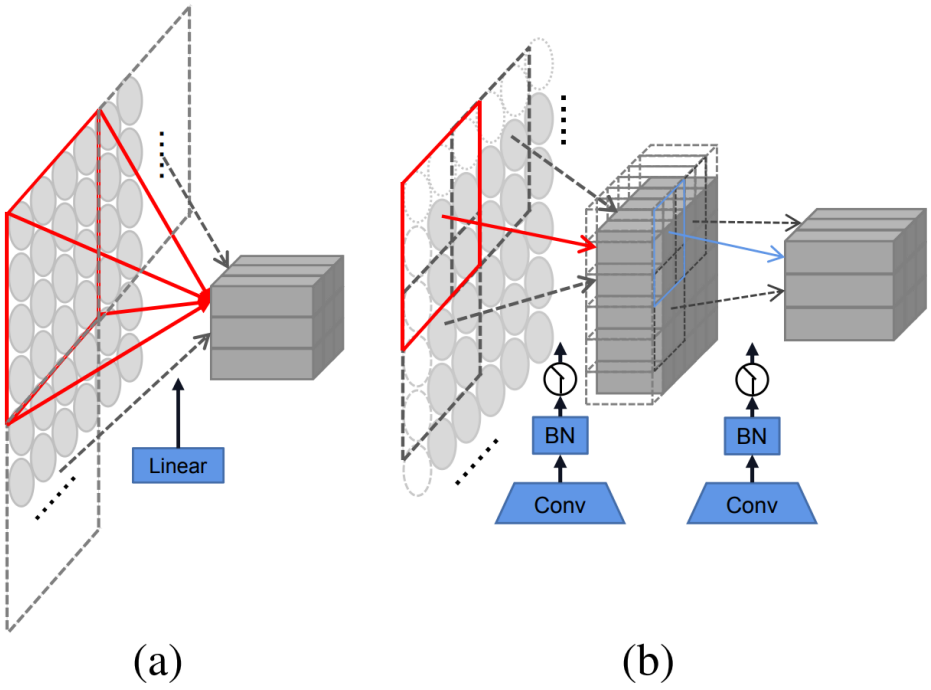
Pregled SVTR modela je prikazan na (Slika 3) i predstavlja tro-faznu mrežu sa progresivno smanjujućom visinom, namenjenu za prepoznavanje teksta. Slika teksta veličine H×W×3, prvo se transformiše u $\frac{H}{4} \times \frac{W}{4}$ isečaka dimenzije D_0 koristeći progresivno preklapajuće ugrađivanje isečaka. Isečci predstavljaju karakterne(znakovne) komponente, od kojih svaka odgovara delu tekstualnog karaktera na slici. Zatim se izvode tri faze, od kojih se svaka sastoji od niza blokova za mešanje praćenih operacijom spajanja ili kombinovanja, na različitim skalama za ekstrakciju karakteristika. Osmišljeni su lokalni i globalni blokovi za mešanje za ekstrakciju lokalnih obrazaca nalik potezima i hvatanje međukomponentne zavisnosti. Pomoću osnovne strukture se karakterizuju komponentne karakteristike i zavisnosti na različitim udaljenostima i na više skala, predstavljene kao C veličine $1 \times \frac{W}{4} \times D_3$, koje percipira karakteristike znakova na više nivoa granularnosti. Na kraju procesa, model istovremeno predviđa sve znakove sa ulazne slike i primenjuje postupak uklanjanja duplikata kako bi se eliminisali eventualno pogrešno ponovljeni karakteri koje je model predvideo, a koji nisu stvarno prisutni na originalnoj slici. Rezultat ovog procesa je konačan niz prepoznatih znakova.



Slika 3: Arhitektura SVTR-a: Mreža koja kroz tri faze progresivno smanjuje visinu mape karakteristika. U svakoj fazi se izvodi niz blokova za mešanje, nakon čega sledi operacija spajanja ili kombinovanja. Na kraju se prepoznavanje vrši linearnim predviđanjem.

2.4.2 Progresivno preklapajuće ugrađivanje isečaka

Pvri korak u obradi slike teksta je njeno razlaganje na manje delove koje nazivamo isečcima. Dobijanje karakterističnih isečaka koji predstavljaju komponente znakova znači prelazak iz $X \in \mathbb{R}^{H \times W \times D_0}$ u $CC_0 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times D_0}$. Postoje dva uobičajena načina da se ovo uradi — korišćenje 4×4 mreže za podelu slike i linearna transformacija svakog dela(Slika 4(a)) i korišćenje 7×7 konvolucionog filtera sa korakom 4. Autori SVTRa su izabrali alternativni metod. Oni koriste dva manja konvoluciona filtera 3×3 jedan za drugim sa korakom 2, kao što je prikazano na (Slika 4(b)). Takođe koriste tehniku zvanu normalizacija serije da bi održali brojeve pod kontrolom. Ovaj novi metod zahteva nešto više računarske snage, ali je bolji u kombinovanju karakteristika iz slike.



Slika 4: (a) Linearna projekcija u ViT [Dos+21]. (b) SVTR progresivno preklapajuće ugrađivanje isečaka.

2.4.3 Blok mešanja

S obzirom na to da se dva karaktera mogu blago razlikovati važno je posmatrati male delove koji čine karaktere. Prepoznavanje teksta se u velikoj meri oslanja na ekstrakciji karakteristika na nivou komponenti karaktera. Međutim, postojeće studije uglavnom koriste niz karakteristika za predstavljanje teksta na slici. Svaka karakteristika odgovara deliću regiona slike, koji je često nerazumljiv, posebno za nepravilan tekst — što nije optimalno za opisivanje karaktera. Nedavni napredak u vizuelnim transformatorima uvodi 2D reprezentaciju karakteristika, ali njeno korišćenje u kontekstu prepoznavanja teksta je još uvek u fazi istraživanja. Autori SVTRa sugerišu da su dve vrste karakteristika važne za prepoznavanje teksta. Prva su lokalni obrasci, kao što su mali detalji koji čine karakter, poput poteza. Oni pokazuju kako su različiti delovi karaktera međusobno povezani i stvaraju se morfološke karakteristike i korelacije između različitih delova karaktera. Druga su međukarakterne zavisnosti, koje se odnose na to kako su različiti karakteri povezani jedni s drugima, ili kako se tekst odnosi na netekstualne delove slike. Da bi uhvatili ove karakteristike, autori su kreirali dva posebna bloka mešanja. Ovi blokovi koriste tehniku zvanu samopažnja, koja pomaže modelu da se fokusira na važne delove slike. Koristeći dva različita područja fokusa koja mehanizam samopažnje razmatra, ovi blokovi mogu uhvatiti i male detalje i širu sliku o tome kako su karakteri međusobno povezani.

Globalno mešanje Kao što se vidi na (Slici 4(a)), globalno mešanje procenjuje zavisnost među svim komponentama karaktera. S obzirom da su tekstualni i ne-tekstualni sadržaj dva glavna elementa na slici, takvo generalno mešanje može uspostaviti dugoročnu zavisnost među komponentama različitih karaktera. Pored toga, ono je takođe sposobno da oslabi uticaj ne-tekstualnih komponenti, istovremeno pojačavajući važnost tekstualnih komponenti. Matematički, za komponente karaktera CC_{i-1} iz prethodne faze, prvo se vrši njihovo preoblikovanje u niz karakteristika. Pri uvođenju u blok mešanja, primenjuje se normalizacija sloja, a zatim se koristi multi-head samopažnja za modelovanje zavisnosti. Nakon toga, sekvencijalno se primenjuju normalizacija sloja i MLP za fuziju karakteristika. Zajedno sa prečicama, formira se blok globalnog mešanja.

Lokalno mešanje Kao što se vidi na (Slici 4(b)), lokalno mešanje procenjuje korelaciju među komponentama unutar unapred definisanog prozora. Njegov cilj je da kodira morfološke karakteristike karaktera i uspostavi veze između komponenti unutar jednog karaktera, što simulira karakteristiku nalik potezu koja je vitalna za identifikaciju karaktera. Za razliku od glo-

balnog mešanja, lokalno mešanje razmatra okolinu za svaku komponentu. Slično konvoluciji, mešanje se odvija koristeći pristup klizećeg prozora. Veličina prozora je empirijski postavljena na 7×11. U poređenju sa globalnim mešanjem, lokalno implementira mehanizam samopažnje za detekciju lokalnih obrazaca. Kao što je prethodno pomenuto, dva bloka mešanja imaju za cilj izvlačenje različitih karakteristika koje su komplementarne. U SVTR-u, blokovi se rekurentno primenjuju više puta u svakoj fazi za sveobuhvatnu ekstrakciju karakteristika.

2.4.4 Spajanje

Održavanje konstantne prostorne rezolucije kroz faze je računski skupo, što takođe dovodi i do redudantnosti reprezentacije karakteristika kroz slojeve. Kao posledica toga, autori SVTR-a osmišljavaju operaciju spajanja nakon blokova mešanja u svakoj fazi (osim u poslednjoj). Karakteristikama koje su izlaz iz poslednjeg bloka mešanja, se prvo menja dimenzija u veličinu $h \times w \times d_{i-1}$, gde h, w i d_{i-1} označavaju trenutnu visinu, širinu i broj kanala, redom. Zatim se primenjuje konvolucija veličine 3×3 sa korakom 2 u dimenziji visine i korakom 1 u dimenziji širine, praćenu normalizacijom sloja, generišući novi sloj dimenzije $\frac{h}{2} \times w \times d_i$.

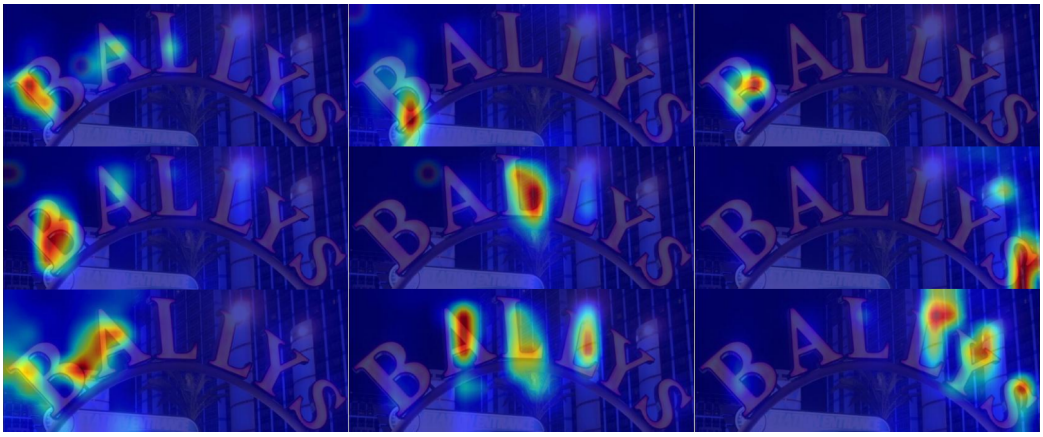
Operacija spajanja prepolovljava visinu dok zadržava konstantnu širinu. Ovo ne samo da smanjuje vremenske troškove obrade, već takođe gradi hijerarhijsku strukturu prilagođenu tekstu. Tipično, većina tekstova na slikama se pojavljuje horizontalno ili blizu horizontalno. Kompresijom dimenzije visine i dalje ostaje uspostavljena višeskalarna reprezentacija svakog karaktera, a pritom ne utiče na raspored isečaka u dimenziji širine. Stoga, smanjivanje dimenzije visine ne povećava šanse za kodiranje susednih isečaka u istu komponentu kroz faze. Takođe, povećava se dimenzija kanala d_i kako bi se nadoknadio gubitak informacija.

2.4.5 Kombinovanje i Predikcija

U poslednjoj fazi, operacija spajanja se zamenjuje operacijom kombinovanja. Prvo se dimenzija visine svede na 1, a zatim se primenjuje potpuno povezani sloj, nelinearna aktivacija i dropout. Na taj način, komponente karaktera se dodatno kompresuju u sekvencu karakteristika, gde je svaki element predstavljen karakteristikom dužine D_3. U poređenju sa operacijom spajanja, operacija kombinovanja može da izbegne primenu konvolucije na slojevima čija je veličina veoma mala u jednoj dimenziji, npr. ukoliko je dimenzija visine 2.

Sa kombinovanim karakteristikama, implementirano je prepoznavanje teksta koristeći jednostavne paralelne linearne predikcije. Konkretno, koristi se linearni klasifikator sa N čvorova. On generiše transkripcijsku sekvencu veličine $\frac{W}{4}$, gde idealno, komponente istog karaktera bivaju transkribovane kao duplikati karaktera, a komponente ne-teksta se transkribuju u prazan simbol. Sekvenca se automatski kondenzuje u konačni rezultat. U implementaciji, N je postavljen na 37 za engleski jezik i 6625 za kineski jezik.

2.4.6 Analiza Vizualizacije



Slika 5: Vizualizacija SVTR mapi pažnje

Svaka mapa se može objasniti kao da ima različitu ulogu u celokupnom prepoznavanju. Ilustracija devet primera mapa je prikazana na (Slici 5). Prvi red prikazuje tri mape koje se fokusiraju na deo karaktera “B”, sa naglaskom na njegovu levu stranu, donji deo i srednji deo, redom. Te tri mape ukazuju na to da različiti regioni karaktera doprinose njegovom prepoznavanju. Drugi red prikazuje tri mape koje se fokusiraju na različite karaktere, tj. “B”, “L” i “S”. SVTR takođe može da nauči karakteristike karaktera posmatrajući karakter kao celinu. Treći red prikazuje tri mape koje istovremeno aktiviraju više karaktera, što implicira da su zavisnosti među različitim karakterima uspešno uhvaćene. Ova tri reda zajedno otkrivaju da prepoznavač hvata tragove na nivou dela karaktera, celog karaktera i više karaktera, u skladu sa tvrdnjom da SVTR percipira višeslojne karakteristike komponenti karaktera, potvrđujući efikasnost SVTR-a.

3 Primene

4 Prikupljanje i rad sa podacima

4.1 Upoznavanje sa podacima

4.2 Augmentacija podataka

4.3 Razvrstavanje i čišćenje podataka

4.4 Pravljenje sintetičkog data seta

4.4.1 Generisanje pozadina tablica

4.4.2 Generisanje teskta na tablicama

5 Implementacija

5.1 Treniranje modela prepoznavanja teksta

5.2 Komponente sistema

5.2.1 Detektor tablica

5.2.2 Detektor teksta

5.2.3 Prepoznavanje teksta

5.3 Tehnologije

5.3.1 PaddlePaddle

5.3.2 Docker

5.3.3 FastAPI

6 Rezultati

7 Buduća poboljšanja

8 Zaključak

Literatura

[SBY15] Baoguang Shi, Xiang Bai i Cong Yao. *An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*. 2015. arXiv: 1507.05717 [cs.CV].

[Li+19] Hui Li i dr. *Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition*. 2019. arXiv: 1811.00751 [cs.CV].

[SCX19] Fenfen Sheng, Zhineng Chen i Bo Xu. *NRTR: A No-Recurrence Sequence-to-Sequence Model For Scene Text Recognition*. 2019. arXiv: 1806.00926 [cs.CV].

[Hu+20] Wenyang Hu i dr. „GTC: Guided Training of CTC towards Efficient and Accurate Scene Text Recognition". U: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (apr. 2020.), str. 11005–11012. DOI: 10.1609/aaai.v34i07.6735. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6735.

[Yu+20] Deli Yu i dr. *Towards Accurate Scene Text Recognition with Semantic Reasoning Networks*. 2020. arXiv: 2003.12294 [cs.CV]. URL: https://arxiv.org/abs/2003.12294.

[Dos+21] Alexey Dosovitskiy i dr. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[Fan+21] Shancheng Fang i dr. *Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition*. 2021. arXiv: 2103.06495 [cs.CV]. URL: https://arxiv.org/abs/2103.06495.

[Liu+21] Ze Liu i dr. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV]. URL: https://arxiv.org/abs/2103.14030.

[Du+22] Yongkun Du i dr. *SVTR: Scene Text Recognition with a Single Visual Model*. 2022. arXiv: 2205.00159 [cs.CV]. URL: https://arxiv.org/abs/2205.00159.

[Zhe+23] Tianlun Zheng i dr. *CDistNet: Perceiving Multi-Domain Character Distance for Robust Text Recognition*. 2023. arXiv: 2111.11011 [cs.CV].