SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

Projektni prijedlog za kolegij Strojno učenje

Detekcija raka na temelju histoloških preparata

Bjanka Bašić

Ivan Knezić

Jelena Lončar

Zagreb, travanj 2019.

Sadržaj

1. Uvodni opis problema 1

2. Cilj i hipoteze istraživanja problema 1

3. Pregled dosadašnjih istraživanja 2

4. Materijali, metodologija i plan istraživanja 2

4.1. Način rješavanja problema 2

4.2. Prikupljanje podataka 2

4.3. Metode/algoritmi/tehnike/alati 2

4.4. Ocjena uspješnosti rezultata projekta 3

5. Očekivani rezultati predloženog projekta 3

6. Popis literature 4

1. Uvodni opis problema

Problem koji nastojimo riješiti problem je detekcije prisutnosti metastatskog raka na temelju digitalnih fotografija histoloških preparata tkiva limfnih čvorova. Taj se problem ustvari svodi na problem binarne klasifikacije fotografija. Radi se o mikroskopskim snimkama tkiva limfnih čvorova, na koja je primijenjeno standardno hematoksilin-eozin (HE) bojenje. Ta je metoda bojenja jedna od najraširenijih u medicinskoj dijagnostici te proizvodi plave, ljubičaste i crvene boje. Tamnoplavi hematoksilin veže se za negativno nabijene tvari, poput nukleinskih kiselina, a ružičasti eozin za pozitivno nabijene tvari, poput aminokiselinskih lanaca (uglavnom proteina). Tipično se jezgre stanica boje plavo, a citoplazma i unutarstanični dijelovi u različite nijanse ružičaste. Histološka procjena metastaza limfnih čvorova dio je određivanja stadija raka dojke. Dijagnostička procedura dugotrajna je i naporna za patologe, budući da je potrebno ispitati veliko područje tkiva, a malene se metastaze lako mogu previdjeti.

Podaci koje koristimo sastoje se od velikog broja malih fotografija koje treba klasificirati, kao i od datoteke koja sadrži stvarne vrijednosti ciljne varijable za pojedine primjere iz *train* skupa. Fotografije u skupu podataka dimenzija su 96 x 96px, a ono što rješavanje ovog problema čini osobito izazovnim jest to što metastaze mogu biti reda veličine jedne stanice usred velikog područja tkiva. Određen se primjer smatra pozitivnim primjerom ako centralno područje fotografije dimenzija 32 x 32px sadrži barem jedan piksel tumorskog tkiva. Tumorsko tkivo u okolnom području fotografije ne utječe na vrijednost ciljne varijable. Skup podataka kombinacija je dvaju nezavisnih skupova podataka prikupljenih u Radboud University Medical Center (Nijmegen, Nizozemska) i University Medical Center Utrecht (Utrecht, Nizozemska). Fotografije su nastale uslijed rutinskih kliničkih postupaka te bi školovani patolog na temelju sličnih preparata nastojao identificirati metastaze. Doduše, moguće je da određen dio relevantnih informacija o okolini nedostaje na fotografijama malih dimenzija našega skupa podataka.

2. Cilj i hipoteze istraživanja problema

Cilj je istraživanja problema detekcija metastatskog raka na temelju fotografija histoloških preparata tkiva limfnih čvorova. U tu svrhu nastojimo naučiti što uspješniji binarni klasifikator, odnosno model koji što uspješnije klasificira fotografije iz prostora primjera χ. Drugim riječima, nastojimo primjenom adekvatnih algoritama strojnog učenja postići da model bude što bolji prediktor za vrijednost ciljne varijable odgovarajućeg elementa prostora primjera. Vrijednost ciljne varijable zapravo je oznaka jedne od dviju mogućih klasa, od kojih jedna označava prisutnost tumorskog tkiva, a druga odsutstvo tumorskog tkiva. Model evaluiramo, tj. njegovu uspješnost određujemo na temelju tehnika opisanih u odsječku 4.1. *Ocjena uspješnosti rezultata projekta*. Hipoteza je istraživanja da je moguće detektirati metastatski rak na temelju fotografija histoloških preparata tkiva limfnih čvorova, odnosno da je moguće primjenom adekvatnih algoritama strojnog učenja dobiti model koji uspješno klasificira fotografije.

3. Pregled dosadašnjih istraživanja

Postojeće metode kojima se problem rješavao uglavnom pripadaju domeni dubokog učenja. Najčešće su korištene neuronske mreže, i to konvolucijske neuronske mreže raznovrsnih arhitektura. Dosadašnja su istraživanja koja su koristila konvolucijske neuronske mreže pokazala njihovu uspješnost u rješavanju našeg problema binarne klasifikacije. Uglavnom su korišteni *pre-trained* modeli, poput DenseNet inačica DenseNet-169 ili DenseNet-121, *pre-trained* modela za PyTorch, optimiziranu tenzor biblioteku za duboko učenje. Međutim, mnoga su dosadašnja istraživanja koristila samo manje dijelove skupa podataka, dok je u našemu planu iskoristiti sve dostupne podatke.

4. Materijali, metodologija i plan istraživanja

4.1. Način rješavanja problema

Problem ćemo rješavati primjenom metoda i algoritama nadziranog učenja. Koristit ćemo neuronske mreže (duboko učenje) i nastojati dobiti što uspješniji model istraživanjem raznolikih mogućnosti za arhitekture neuronskih mreža te hiperparametre.

4.2. Prikupljanje podataka

Podatke preuzimamo s web stranice **Kaggle**. Radi se o skupu podataka koji je blago modificirana verzija PatchCamelyon (PCam) benchmark skupa podataka, dostupnog na sljedećem linku: **https://github.com/basveeling/pcam**. Originalni PCam skup podataka sadrži duplicirane fotografije zbog slučajnog uzorkovanja, dok verzija dostupna na Kaggleu ne sadrži duplikate.

4.3. Metode/algoritmi/tehnike/alati

U svrhu rješavanja problema planiramo koristiti programski jezik Python te raznovrsne biblioteke za njega, poput Numpy, Pandas, Scikit-learn, Keras, koji je dizajniran kako bi omogućio brzo eksperimentiranje s dubokim neuronskim mrežama te Fast.ai, koji je izgrađen na temelju PyTorch. Kao razvojno okruženje koristit ćemo Jupyter. Problem ćemo rješavati metodama i algoritmima dubokog učenja, uporabom neuronskih mreža. Planiramo koristiti konvolucijske neuronske mreže i istražiti pomoću kojih je njihovih arhitektura te hiperparametara (primjerice, nastojat ćemo pronaći optimalan *learning rate* za *gradient descent*) moguće dobiti najuspješniji model. Također, u svrhu poboljšanja sposobnosti modela da prepozna različite verzije fotografije, osiguravanja da je model robusniji na blage promjene u podacima i povećanja opsega informacija kojima model raspolaže, na podatke (fotografije) planiramo primijeniti regularizacijsku tehniku koja se sastoji od augmentacijskih transformacija. Kao rezultat očekujemo dobiti model koji je sposobniji prepoznati ciljane objekte na fotografijama varirajućeg kontrasta, veličine, snimljenih iz različitih kuteva i slično. Neke od transformacije koje planiramo primjenjivati uključuju: osnovne transformacije (rotiranje te promjene osvjetljenja), *side-on* transformacije (uz rotiranje te promjene osvjetljenja i zrcaljenje u odnosu na vertikalnu os) te *top-down* transformacije (uz rotiranje te

promjene osvjetljenja i zrcaljenje u odnosu na horizontalnu os). Slično, planiramo *test* fotografije modificirati pomoću Test Time Augmentation (TTA).

4.4. Ocjena uspješnosti rezultata projekta

Model ćemo primarno evaluirati pomoću vrijednosti AUC (Area Under Curve), odnosno površine ispod ROC (Receiver Operating Characteristic) krivulje, krivulje koja prikazuje odnos TPR (true positive rate) u odnosu na FPR (false positive rate). Pritom je TPR broj korektnih klasifikacija u (pozitivnoj) klasi u odnosu na ukupan broj pozitivnih primjera (TPR = tp/(tp + fn)), dok je FPR definiran kao broj krivih klasifikacija u (pozitivnoj) klasi u odnosu na ukupan broj negativnih primjera (FPR = fp/(fp + tn)). Težnja je ostvariti što veću vrijednost AUC, približiti se vrijednosti savršenog klasifikatora (AUC = 1). S obzirom na prirodu problema, odnosno činjenicu da se radi o problemu iz domene medicinske dijagnostike, pozitivni su nam primjeri daleko važniji od negativnih primjera pa uspješnost modela planiramo evaluirati i određivanjem konkretne vrijednosti osjetljivosti/recall/true positive rate. Također, odredit ćemo i matricu konfuzije, vrijednost preciznosti (P = tp/tp + fp)) te F1, mjere koja povezuje preciznost i osjetljivost.

5. Očekivani rezultati predloženog projekta

Kao konačni rezultat projekta nastojat ćemo predati model koji što uspješnije rješava dani binarni klasifikacijski problem, odnosno model sa (primarno) što većom vrijednošću AUC. Očekujemo postići AUC vrijednost iznad 0.9.

6. Popis literature

[1] Chollet, F. (2017) *Deep Learning with Python*. 1. izd. SAD: Manning Publications

[2] *CS231n: Convolutional Neural Networks for Visual Recognition* [online]. Stanford CS class. Dostupno na: **https://cs231n.github.io/convolutional-networks/**

[3] Zulkifi, H. (2018) *Understanding Learning Rates and How It Improves Performance in Deep Learning* [online]. Towards Data Science. Dostupno na: **https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10**

[4] Godoy, D. (2018) *Understanding binary cross-entropy / log loss: a visual explanation* [online]. Towards Data Science. Dostupno na: **https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a**

[5] Doshi, N. (2018) *Augmentation for Image Classification* [online]. Towards Data Science. Dostupno na: **https://towardsdatascience.com/augmentation-for-image-classification-24ffcbc38833**

[6] Brownlee, J. (2019) *How to Configure the Learning Rate Hyperparameter When Training Deep Learning Neural Networks* [online]. Machine Learning Mastery. Dostupno na: **https://machinelearningmastery.com/learning-rate-for-deep-learning-neural-networks/**

[7] Gugger, S. (2018) *The 1cycle policy* [online]. Dostupno na: **https://sgugger.github.io/the-1cycle-policy.htmlhttps://sgugger.github.io/the-1cycle-policy.html**

[8] Ioffe, S. i Szegedy, C. (2015) *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* [online]. Cornell University. Dostupno na: **https://arxiv.org/abs/1502.03167**

[9] Šmuc, T. (2018/2019) Predavanja [online]. Dostupno na: **https://web.math.pmf.unizg.hr/nastava/su/materijali/**