# O klasifikaciji slik v ne-enolično določljive razrede

**Jon Natanael Muhovič, Domen Tabernik, Danijel Skočaj**

*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*
*E-pošta: {jon.muhovic,domen.tabernik,danijel.skocaj}@fri.uni-lj.si*

## About classification of images into non-uniquely identifiable classes

*Image classification is one of the most basic and frequently addressed computer vision tasks. The usual formulation of this tasks requires classification of an image into the one of several possible classes. The most common metric for measuring the classifier's performance is classification accuracy, defined as a percentage of correctly classified images. However, such formalisation of the classification problem relies on a strong assumption that for every image a category is uniquely identifiable and assigned by the domain expert. In this paper we address scenarios where this assumption does not hold. In particular, we present an analysis of the results obtained by the convolutional neural network and twelve participants who were tasked to classify the images of planks into eight classes and discuss the label ambiguity problem.*

## 1 Uvod

Klasifikacija slik je ena najbolj osnovnih in uporabljanih funkcionalnosti v računalniškem vidu. Zelo pogosto se namreč srečamo s problemom, ko je potrebno razpoznati kaj je prikazano na sliki, oz. prikazano uvrstiti v primeren razred. Zato se raziskovalci s tem problem ukvarjajo od najzgodnejših dni računalniškega vida naprej.

### 1.1 Definicija problema

Običajno je problem klasifikacije formuliran na zelo enostaven način. Vsebino slike je potrebno klasificirati natanko v enega izmed $n$ razredov. Na podoben način nato merimo tudi uspešnost klasifikatorja. Uveljavljena mera je klasifikacijska točnost, ki predstavlja odstotek slik, ki so bile pravilno klasificirane. Včasih to zahtevo celo malce omilimo in smatramo klasifikacijo kot uspešno, če klasifikator vrne pravilni razred med prvimi $k$ napovedmi. Pri tem seveda tudi predpostavimo, da je kategorija za vsako sliko oz. njeno vsebino enolično določljiva ter pravilno določena s strani poznavalca problema. Celoten postopek učenja klasifikatorja ter njegove evalvacije sloni na tej predpostavki.

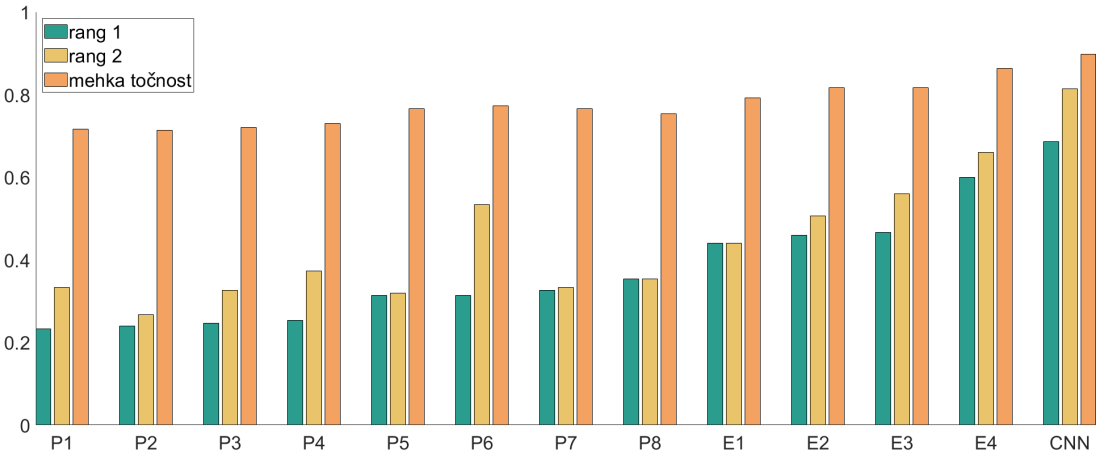
Pri realnih problemih pa ta predpostavka ne drži vedno. Nekatere kategorije so dvoumne, nekatere vsebujejo druge kategorije, nekatere slike se lahko uvrstijo v več kategorij ali pa različni uporabniki uvrstijo isto sliko v različne kategorije. To predstavlja pri klasičnih pristopih učenja in evalvacije dodaten izziv. V tem članku bomo obravnavali problem ne-enolično določljivih razredov na primeru industrijske aplikacije klasifikacije desk. Naloga bo klasificirati deske v enega izmed osmih kakovostnih razredov. Kot je razvidno s Slike 1, je izgled desk, ki pripadajo posameznim razredom, dokaj podoben. Razredi se med seboj razlikujejo v podrobnostih, kot sta odtenek in homogenost barve ter velikost in pogostost grč. Meje med razredi je zelo težko enolično postaviti. Zato tudi ni vseeno kakšne napake sistem dela. Včasih je popolnoma sprejemljivo, če je slika klasificirana v enega izmed dveh sosednjih razredov, medtem ko klasifikacija v bolj različen razred prestavlja veliko večjo napako.

Slika 1: Predstavniki osmih razredov desk.

V tem članku bomo obravnavali ta problem. Analizirali bomo rezultate, ki smo jih dobili s klasifikatorjem, ki temelji na globokem učenju, in rezultate klasifikacij dvanajstih posameznikov, tako poznavalcev dotične aplikativne domene, kot tudi laikov, ter jih analizirali z vidika ne-enolične določljivosti razredov.

### 1.2 Sorodna dela

Ne-enolična določljivost razredov se v literaturi pogosto naslavlja z uporabo mehkih označb razredov [5] (ang. soft labels) ali z glajenjem označb razredov [6] (ang. label smoothing). V obeh primerih enični vektor (ang. one-hot vector), s katerim označujemo kategorijo, zamenja drugačna označba, in sicer mehke označbe preko več kategorij oz. verjetnostna porazdelitev uvrstitve v posamezne razrede, ki odraža podobnost oz. ne-enoličnost kategorij. V primeru glajenja razredov se običajno uporabi enakomerna porazdelitev med razredi, ki izboljša klasifikacijsko točnost. Kot so pokazali Müller in ostali [4], to povzroči zoženje posamezne gruče razredov v podprostoru značilk, vendar pa to ni povsem zaželeno v prime-




Slika 2: Rezultati klasifikacij.

rih, ko se ne moramo zanašati na točnost podane referenčne označbe oz. zlatega standarda. V takih primerih se za bolj primerne izkažejo mehki razredi, kjer se enični vektor zamenja s porazdelitvijo, ki pravilno odraža nedoločenost med posameznimi razredi. Ta porazdelitev se nato uporabi v razširjeni kriterijiski funkciji prečne entropije [5], ki skuša klasifikator zgraditi na način, da se njegove napovedi čim bolj ujemajo distribucijami pravilnih napovedi posameznih slik. Pristop se pogosto uporablja pri klasifikaciji splošnih objektov, kjer je mogoče nedoločenost med kategorijami definirati na podlagi semantične hierarhije [1].

Ti pristopi se torej osredotočajo na mehkejše napovedovanje razredov, ki ustreza dvoumnosti med razredi v skladu z dvoumnostjo napovedi ljudi. Problem, s katerim se ukvarjamo v tem članku, pa je rahlo drugačen. Osredotočamo se namreč na industrijsko aplikacijo, kjer je pomembna zgolj ena napoved, ki se upošteva v nadaljnjem proizvodnem procesu, in ne celotna distribucija napovedi. Se pa sorodnost med posameznimi razredi izraža pri evalvaciji napovedi. S stališča proizvodnega procesa je namreč veliko manj izrazita napaka uvrstitve slike v soroden razred, torej v razred, ki ima podano dokaj visoko (čeprav ne najvišjo) verjetnost, kot v razred, kjer je ta verjetnost, oz. ustreznost, majhna.

## 2 Metode

Glavni namen te raziskave je raziskati vpliv dvoumnosti pri določevanju razredov slik na rezultate klasifikacije. Zato smo izvedli študijo z 12 udeleženci in rezultate primerjali z napovedmi avtomatskega klasifikatorja.

Za eksperimentalno evalvacijo smo pridobili osem laikov in štiri eksperte. Vsak od njih je moral za vsako od slik v zbirki določiti razred. Zbirka je obsegala 150 primerov desk, od tega jih je bilo 100 naključno izbranih iz celotne zbirke slik desk, preostalih 50 pa je bilo naključno izbranih iz prvih 100 in naključno premešanih. Celotna zbirka je tako vsebovala eno tretjino ponovljenih slik, oz. 50 parov duplikatov, na kar udeleženci niso bili opozorjeni.

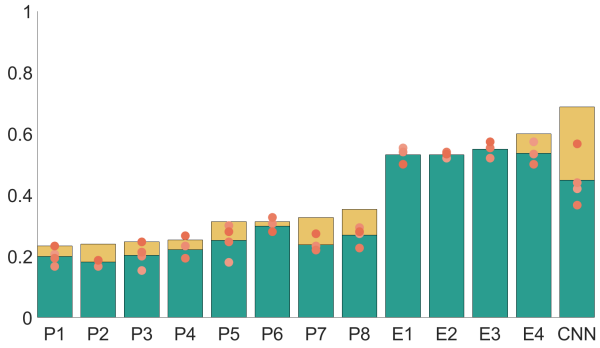
Udeleženci so za pomoč pri klasifikaciji dobili 14 primerov desk iz vsakega razreda. Za vsako od slik v eksperimentalni zbirki so morali udeleženci določiti razred, za katerega so menili, da mu posamezna slika pripada. V primeru negotovosti so lahko določili še drugi razred, za katerega so menili, da bi mu posamezna slika lahko pripadala.

Za avtomatsko klasifikacijo slik smo uporabili arhitekturo ResNet50 [3], prednaučeno na zbirki ImageNet [2]. Mrežo smo učili za klasifikacijo vsakega od osmih razredov desk, za učenje pa smo uporabili našo podatkovno zbirko z več kot 2500 slikami, ki ni vsebovala slik, ki so bili uporabljene v evalvacijski študiji. Za učenje klasifikacije smo uporabili standardno kategorično križno entropijo in optimizator ADAM.

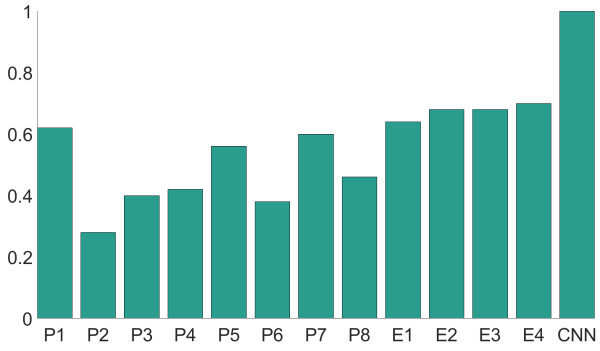
**Metrike:** Uspešnost rezultatov smo merili s štirimi metrikami:

(i) *rang 1*, ki meri klasifikacijsko točnost s prvo napovedjo,
(ii) *rang 2*, ki meri uspešnost klasifikacije s prvo ali z drugo napovedjo,
(iii) *mehka točnost*, kjer se nepravilni odgovori obtežijo v odvisnosti od kritičnosti napake; pri tem se klasifikacija v sosednje razrede veliko manj kaznuje kot klasifikacija v zelo različne razrede,
(iv) *ponovljivost*, ki izračuna odstotek istih testnih primerov, ki so bili obakrat enako klasificirani.

Poleg standardne klasifikacijske točnosti, ki se običajno uporablja pri evalvaciji klasifikacijskih rešitev, smo torej uporabili tudi mehko klasifikacijsko točnost. Ta bolj realno zajame uspešnost metode, saj ne obravnava vseh napak kot enako problematične. To je v praksi pogosto potrebno, saj vse napake nimajo enake teže, kar predpostavlja klasična klasifikacijska točnost. Težo napake uravnavamo z vpeljano matriko podobnosti, ki napačni klasifikaciji pripiše ne-ničelno ceno, glede na podobnost pravilnega razreda. Mehko klasifikacijsko točnost tako računamo kot $SCA = \frac{1}{N}\sum_{i=1}^{N} w_{c_i,y_i}$, kjer je $N$ število vseh slik, $w_{a,b}$ pa predstavlja utež za primer, ko je napoved klasifikacije slike iz razreda $a$, razred $b$. Ta utež je velika (blizu 1), če sta si razreda podobna in 0, če slika nikakor ne bi smela biti klasificirana v ta razred. V primeru standardne klasifikacijske točnosti je matrika $W$ kar




Slika 3: Rezultati z različnimi zlatimi standardi (rang-1).




Slika 4: Delež ponovljivih odgovorov.

diagonalna matrika. V praksi je ta matrika odvisna od lastnosti in podobnosti posameznih razredov, ki jih določijo eksperti.

## 3 Eksperimentalni rezultati

**Klasifikacijska točnost:** Eksperimentalni rezultati so prikazani na Sliki 2, kjer so s $P1 - P8$ označeni laični anotatorji, medtem ko so z $E1 - E4$ označeni ekspertni anotatorji. Iz eksperimentalnih rezultatov po rangu 1 in 2 vidimo, da je točnost laičnih anotatorjev precej nizka, saj so si slike med sabo precej podobne, predvsem tiste iz višjih kvalitetnih razredov. Najvišja točnost laičnih anotatorjev je tako 0.35 po rangu 1. Ekspertni anotatorji se po pričakovanjih odrežejo precej bolje, s točnostmi med 0.44 in 0.6 po rangu 1. Še vedno pa se njihova točnost ne približa vrednosti 1, kar priča o težavnosti oz. dvoumnosti klasifikaije. Pri tem je potrebno poudariti tudi, da imajo v praksi eksperti več informacij za odločanje kot zgolj eno sliko, zato bi bila njihova uspešnost v praksi seveda veliko višja. Najvišjo točnost doseže nevronska mreža, učena s standardno kategorično križno entropijo.

Pri evalvaciji z mehko točnostjo opazimo, da se tako merjena uspešnost poveča vsem udeležencem, kar ni presenetljivo, saj so v tem primeru napake manj kaznovane. Že sam naključni klasifikator bi v tem primeru dosegel uspešnost 0.48, kar je veliko višja osnova od 0.125 v primeru klasične klasifikacijske točnosti. Vidimo tudi, da se je predvsem dvignila mehka točnost laikov, kar pomeni, da so ti pogosto zgrešili, a da napake, ki so jih naredili, niso bile tako problematične.

**Ponovljivost rezultatov:** V eksperimentalno zbirko smo ponovljene slike vključili, da bi preverili konsistentnost človeške klasifikacije. Deske nekaterih razredov so na pogled zelo podobne, kar lahko povzroča težave tudi ekspertom. Na Sliki 4 je prikazan odstotek ponovljenih slik, za katere je posameznik obakrat izbral isti razred.

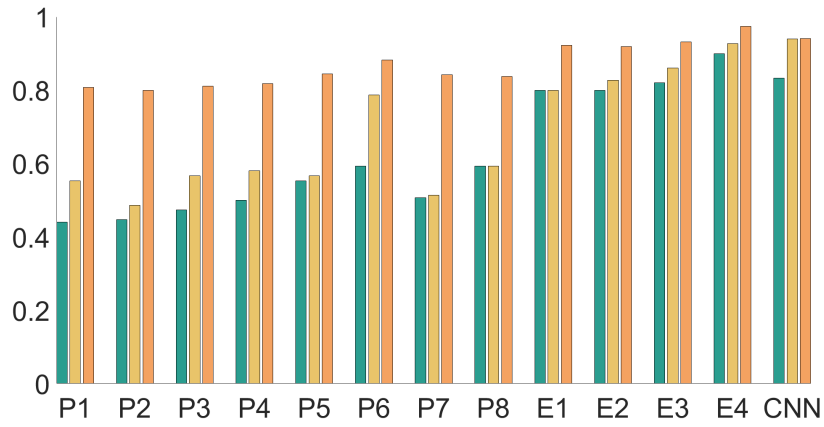
Rezultati prikazujejo precej nizko ponovljivost. Med laiki ponovljivost močno variira, pri čemer sta $P2$ in $P6$ enako klasificirala le 30% oz. 40% primerov, kar dejansko priča o kompleksnosti problema. Po drugi strani so bili eksperti pri klasifikaciji bolj ponovljivi, vendar še vedno ne presegajo 70% ponovljivosti. Model globokega učenja seveda dosega 100% ponovljivost, saj pri detekciji niso prisotni nobeni stohastični procesi.

**Različni eksperti kot zlati standard** Ker so se pri klasifikaciji pogosto razlikovala tudi mnenja ekspertov, smo naredili še evalvacijo pri kateri smo za zlati standard ("pravilno" vrednost razreda za posamezno sliko) vzeli napovedi posameznih ekspertov. Na Sliki 3 prikazujemo za vsakega udeleženca s štirimi dodanimi pikami klasifikacijsko točnost odgovorov pri zlatih standardih štirih ekspertov. Z zelenim stolpcem prikazujemo povprečje teh štirih vrednosti, medtem ko z rumeno barvo prikazujemo razliko do točnosti pri dejanskem zlatem standardu (kot na Sliki 2). Pri ekspertih vzamemo za zlati standard le preostale tri eksperte.

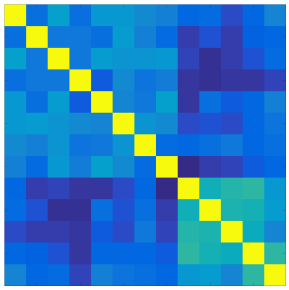
V primeru ko vzamemo druge eksperte za zlati standard se pri večini udeležencev rezultat lahko močno razlikuje med enim in drugim zlatim standardom. Pri večini udeležencev je variacija z različnimi zlatimi standardi lahko tudi po 10%. To implicira veliko razliko med odgovori različnih ekspertov, posledično pa vpliva tudi na učenje globoke metode. Pri slednji lahko opazimo veliko varianco med točnostjo ob različnih zlatih standardih na eni strani, ter izrazito zmanjšanje točnosti, ko uporabimo druge zlate standarde na drugi strani. To nakazuje, da se je globoka metoda naučila specifičnosti podanega zlatega standarda pri učenju, ki pa pogosto ne odraža podobnega zlatega standarda kot ga dojemajo preostali eksperti.

Zmanjšanje povprečne točnosti pri uporabi različnih zlatih standardov je opazno tudi pri laičnih udeležencih ($P1 - P8$), medtem ko tega ni mogoče zaznati pri ekspertih ($E1 - E3$). To nakazuje, da so v povprečju eksperti bolj usklajeni kot laični udeleženci. Razlika je opazna le pri najboljšem ekspertu $E4$, saj je tudi osnovni zlati standard modeliran po istem ekspertu.

**Množica ekspertov kot zlati standard** Kot zlati standard smo dodatno evalvirali tudi množico ekspertov. Na Sliki 5a prikazujemo rezultate, kjer smo za pravilni odgovor šteli tisti razred, ki ga je navedel vsaj eden izmed ekspertov. Pri $E1 - E4$ smo uporabili le preostale tri eksperte. Ta postopek sloni na predpostavki, da so tisti razredi, ki jih navajajo eksperti, veliko bolj pravilni kot tisti, ki jih niti en ekspert ni navedel. V tem primeru pričakovano rezultati povsod poskočijo. Najbolj očitno je povečanje točnosti po vseh metrikah pri ekspertih, kjer so rezultati poskočili tudi preko 80% celo pri klasifikacij-




(a)



(b)

Slika 5: (a) Rezultati z upoštevanjem mnenja katerega koli izmed ekspertov kot pravilno (top5), ter (b) matrika klasifikacijske točnosti pri različnih zlatih standardih.

ski točnosti ranga 1. To nakazuje, da so napake ki so jih delali eksperti, zelo podobne odgovorom drugih ekspertov.

Ta lastnost se odraža tudi v matriki točnosti na Sliki 5b, ki prikazuje klasifikacijsko točnost posameznega udeleženca, ko je za zlati standard vzeta napoved vsakega drugega udeleženca. Na njej je mogoče videti dva vizualna bloka. Prvi blok zajema vseh osem laikov, medtem ko drugi blok zajema vse štiri eksperte ter CNN. Opazna razlika med blokoma pomeni, da so napake, ki jih delajo laiki, med seboj podobne, in ravno tako, da so napake, ki jih delajo eksperti, med seboj podobne. Model globoke mreže se uvrsti v drugi blok, kar pomeni, da so napake, ki jih dela, tipično zajete v odgovorih drugih ekspertov.

## 4 Zaključek

V članku smo želeli raziskati problem nezanesljivih oz. težko določljivih označb v specifičnih problemskih domenah, kjer niti eksperti ne morejo vedno enolično in konsistentno določiti pravilnih razredov. Na problemu klasifikacije kakovosti desk smo pokazali, da lahko globoko učenje močno preseže rezultate laikov ter doseže rezultate primerljive ekspertom oz. v določenih primerih celo boljše, hkrati pa smo opozorili na problem neenolične določljivosti razredov.

Pokazali smo, da so napake laikov veliko bolj podobne med seboj, kot so podobne napakam ekspertov, ter obratno. Pri tem tudi klasifikator, ki temelji na globoki nevronski mreži, izkazuje napake, ki so bolj podobne napakam ekspertov. Izkaže pa se tudi, da se v primeru, ko za zlati standard vzamemo odgovor katerega koli izmed odgovorov ekspertov, rezultati nevronske mreže ne izboljšajo tako zelo kot rezultati ostalih udeležencev. Tak rezultat je pričakovan, saj globoka metoda zgradi model ob upoštevanju podanega zlatega standarda, pri čemer ne eksplicitno upošteva sorodnosti med razredi, oz. se tega zaveda samo implicitno preko ugotovljene vizualne podobnosti.

Ravno to dejstvo predstavlja potencial za izboljšanje metode za avtomatsko klasifikacijo. V primeru podane matrike uteži napake bi bilo smiselno to informacijo neposredno upoštevati že pri učenju in tako omogočiti metodi, da se nauči bolj izogibati klasifikaciji v razrede, ki so za podani primer bolj kritični. V ta namen smo preizkusili metodo mehkih označb razredov [5], ki ni dosegla zadovoljivih rezultatov. Bolj obetaven je pristop z uteževanjem funkcije izgube, ki ga nameravamo dodelati v našem prihodnjem delu.

Vsekakor pa se je v aplikativnih raziskavah za reševanje konkretnih problemov v industriji potrebo zavedati, da je uspešnost delovanja končne aplikacije zelo odvisna od podanega zlatega standarda, ki mora biti čim bolj verodostojen. V primerih, ko le-ta ni enolično določljiv, pa je potrebno to dejstvo upoštevati tako pri učenju, kot pri evalvaciji tovrstnih klasifikacijskih sistemov.

## Zahvala

To delo je bilo delno financirano s strani ARRS projekta J2-9433 (DIVID) in raziskovalnega programa Računalniški vid (P2-0214). Posebna zahvala gre tudi podjetju Menina d.o.o. za pripravo slikovnih podatkov, ter vsem anotatorjem, tako laičnim kot ekspertnim.

## Literatura

[1] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making Better Mistakes: Leveraging Class Hierarchies with Deep Networks. In *CVPR*, pages 12506–12515, 2020.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[4] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When Does Label Smoothing Help? In *NeurIPS*, 2019.

[5] Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. *ICCV*, Oct 2019.

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, volume 2016-Decem, pages 2818–2826, 2016.