

SVEUČILIŠTE U ZAGREBU  
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 872

# **VARIJACIJSKO UČENJE NA ZAŠUMLJENIM OZNAKAMA**

Dominik Jambrović

Zagreb, lipanj, 2025.

Zagreb, 3. ožujka 2025.

## DIPLOMSKI ZADATAK br. 872

Pristupnik: **Dominik Jambrović (0036534818)**

Studij: Računarstvo

Profil: Računarska znanost

Mentor: prof. dr. sc. Siniša Šegvić

Zadatak: **Varijacijsko učenje na zašumljenim oznakama**

Opis zadatka:

Raspoznavanje slika važan je problem računalnog vida s mnogim zanimljivim primjenama. U posljednje vrijeme stanje tehnike postižu duboki modeli zasnovani na konvolucijama i slojevima pažnje. Međutim, standardni postupci teško se nose sa zašumljenim oznakama. U okviru rada, potrebno je odabrati okvir za automatsku diferencijaciju te upoznati biblioteke za rukovanje tenzorima i slikama. Proučiti i ukratko opisati postojeće duboke arhitekture za raspoznavanje slika s posebnim naglaskom na prednaučene samonadzirane modele. Odabrati slobodno dostupne skupove slika te oblikovati podskupove za učenje, validaciju i testiranje. Formulirati optimizacijski cilj s latentnim predikcijama čistih razreda te predložiti rješenje utemeljeno na varijacijskoj aproksimaciji te maksimiziranju očekivanja. Komentirati učinkovitost učenja i zaključivanja. Predložiti pravce za budući rad. Radu priložiti izvorni i izvršni kod razvijenih postupaka, ispitne slijedove i rezultate, uz potrebna objašnjenja i dokumentaciju. Citirati korištenu literaturu i navesti dobivenu pomoć.

Rok za predaju rada: 4. srpnja 2025.

*Zahvale!*

# Sadržaj

<b>1. Uvod</b>	<b>3</b>
<b>2. Problem zatrovanih podataka</b>	<b>5</b>
2.1. Primjeri metoda trovanja podataka	6
2.1.1. Napad BadNets	6
2.1.2. Napad Blend	7
2.1.3. Napad WaNet	7
2.2. Primjeri algoritama za obranu od zatrovanih podataka	8
2.2.1. Obrana ABL	8
2.2.2. Obrana DBD	9
2.2.3. Obrana ASD	11
<b>3. Problem zašumljenih oznaka</b>	<b>13</b>
3.1. Primjeri metoda zašumljivanja oznaka	14
3.1.1. Simetrično zašumljivanje	14
3.1.2. Asimetrično zašumljivanje	15
3.2. Primjeri algoritama za učenje na zašumljenim oznakama	16
3.2.1. Algoritam SOP	16
3.2.2. Algoritam ILL	18
<b>4. Samonadzirano učenje</b>	<b>20</b>
4.1. Kontrastno samonadzirano učenje	21
4.2. Okvir All4One	22
4.2.1. Kontrastiranje najbližih susjeda	23
4.2.2. Kontrastiranje centroida	24
4.2.3. Kontrastiranje značajki	25

4.2.4. Kombinirano kontrastiranje . . . . .	26
<b>5. Algoritam maksimizacije očekivanja . . . . .</b>	<b>27</b>
5.1. Osnovna formulacija algoritma . . . . .	28
5.2. Općenita formulacija algoritma . . . . .	28
<b>6. Transportni problem . . . . .</b>	<b>29</b>
<b>7. VIBE . . . . .</b>	<b>30</b>
<b>8. Skup podataka . . . . .</b>	<b>31</b>
<b>9. Eksperimenti . . . . .</b>	<b>32</b>
<b>10. Zaključak . . . . .</b>	<b>33</b>
<b>Literatura . . . . .</b>	<b>34</b>
<b>Sažetak . . . . .</b>	<b>39</b>

# 1. Uvod

Duboki modeli koriste se u brojnim aspektima naše svakodnevice. Pri razvoju i učenju modela, pažnju prije svega posvećujemo performansama na neviđenim podacima - želimo naučiti modele koji dobro generaliziraju. Drugim riječima, želimo da modeli daju ispravna predviđanja za viđene, ali i za neviđene podatke. Ovime osiguravamo da naša rješenja imaju primjenu i van laboratorijskih uvjeta u kojima se uče.

U procesu razvoja modela za određeni zadatak strojnog učenja, osim odabira arhitekture, algoritma učenja i hiperparametara, veliku ulogu igraju podatci na kojima učimo. Općenito govoreći, prikupljanje i označavanje podataka jedan je od najskupljih dijelova procesa razvoja rješenja za nekih problem. Važno je da prikupljeni podatci što realističnije predstavljaju stvarne situacije s kojima će se naš model susretati tj. da distribucija podataka odgovara stvarnoj distribuciji situacija koje prikazuju. Dodatno, pokazuje se da duboki modeli uz dovoljan kapacitet mogu naučiti ispravno predviđati oznake čak i za nasumično označene podatke [1], tako da je veoma važno da su prikupljeni podatci što točnije označeni.

Područje računalnog vida [2] bavi se razvojem algoritama i modela za brojne zadatke raspoznavanja i razumijevanja slika. Najčešći zadatak je klasifikacija slika - model na ulazu dobiva sliku, a na izlazu treba predvidjeti razred koji odgovara ulaznom primjeru. Iako postoje brojni skupovi slikovnih podataka koji se mogu koristiti za učenje i evaluaciju modela, za konkretne zadatke u većini slučajeva trebamo prikupiti i označiti vlastite slike. Pritom postoji nekoliko čestih opasnosti: prisutnost zatrovanih podataka [3] ili zašumljenih oznaka [4].

Kada govorimo o trovanju podataka, maliciozni agent u skup podataka dodaje zatrovane podatke s ciljem manipulacije izlaza naučenog modela za određene ulaze. S druge strane, anotator podataka bez zlih namjera određenim podacima može pridijeliti netočne oznake, time dodajući podatke sa zašumljenim oznakama u skup. Kroz vrijeme, razvili su se brojni algoritmi za obranu modela od zatrovanih podataka [5, 6, 7], kao i za učenje na zašumljenim oznakama [8, 9]. Ipak, većina radova se fokusira na samo jedan od ovih problema, a ne na razvoj algoritma koji se može nositi s oba problema.

Cilj ovog rada je reproducirati i poboljšati rezultate okvira za obranu od zatrovanih podataka imena VIBE (engl. *Variational inference for backdoor elimination*) [10]. Osim ovoga, cilj je i primijeniti VIBE na problem zašumljenih oznaka. Pritom VIBE evaluiramo na nekoliko čestih vrsta trovanja odnosno zašumljivanja oznaka kako bi se osigurala robusnost okvira. Dodatno, cilj je usporediti VIBE sa stanjem tehnike (engl. *state of the art* - *SotA*) za problem zašumljenih oznaka.

## 2. Problem zatrovanih podataka

Cilj dodavanja zatrovanih podataka [3] u skup je ugrađivanje stražnjih vrata (engl. *backdoor*) u naučeni model. Ako napad uspije, napadač može kontrolirati izlaz modela koristeći suptilne izmjene ulaznog primjera. Općenito govoreći, stvaranje zatrovanih podataka podrazumijeva dodavanje vizualnog okidača na ulazni primjer, kao i prikladnu izmjenu oznaka. Pritom napadač radi izmjenu određenog udjela podataka, dok preostali podatci ostaju neizmjenjeni. Hiperparametar koji opisuje udio zatrovanih podataka zvat ćemo stopom trovanja (engl. *poisoning rate*). Pojedine metode trovanja podataka razlikuju se po načinu dodavanja okidača tj. načinu izmjene ulaznih primjera, kao i po načinu izmjene oznaka.

Kada govorimo o načinu izmjene ulaznih primjera, možemo napraviti podjelu na lokalne i globalne izmjene primjera. Kod lokalnih izmjena, mijenja se samo određeno područje slike, najčešće dodavanjem zadanog okidača na to područje [11]. S druge strane, kod globalnih izmjena se mijenja cijela slika koristeći različite tehnike poput miješanja slike s okidačem [12] ili transformiranja slike na temelju zadanog deformacijskog polja [13]. Osim korištenja jednog okidača za sve zatrovane podatke, određeni napadi koriste okidače specifične za pojedini uzorak [14].

Većinu napada možemo svrstati u jedan od dva načina izmjena oznaka: *all-to-one* i *all-to-all* izmjena oznaka [15]. Kod *all-to-one* metode, primjeri dobivaju zatrovanu oznaku jednog proizvoljno odabranog razreda neovisno o originalnim oznakama pojedinih primjera. S druge strane, kod *all-to-all* metode, primjeri dobivaju zasebne zatrovane oznake ovisno o originalnim oznakama. Osim ova dva načina izmjena oznaka, određeni napadi uopće ne mijenjaju oznake zatrovanih primjera, već se oslanjaju isključivo na jače izmjene ulaznih primjera. Ovakve napade zovemo napadi s čistim oznakama (engl. *clean-label attacks*) [16].



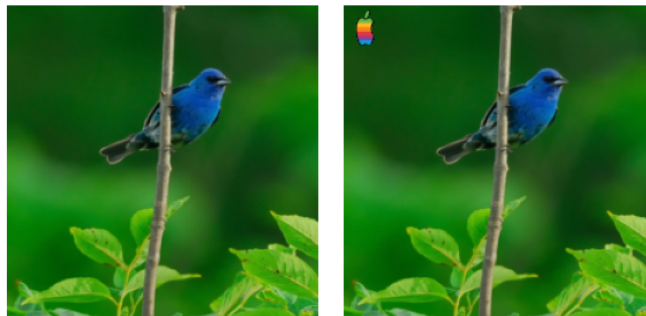
Uspješnost pojedinog napada mjerimo metrikom imena stopa uspješnosti napada (engl. *attack success rate* - ASR). Ovu mjeru definiramo kao točnost modela mjerenu isključivo na zatrovanim primjerima. Cilj algoritama za obranu od zatrovanih podataka je naučiti čisti model na zatrovanom skupu. Drugim riječima, glavni cilj obrane je naučiti model koji ima izvrsne performanse na čistim podacima, ali i što nižu stopu uspješnosti napada.

## 2.1. Primjeri metoda trovanja podataka

U ovome radu, fokusiramo se na tri metode trovanja podataka: napade BadNets [11], Blend [12] te WaNet [13].

### 2.1.1. Napad BadNets

Napad BadNets uobičajeno dodaje jedan zadani okidač na svaki odabrani ulazni primjer. Okidač možemo shvatiti kao uzorak piksela koji se dodaje na specifično mjesto na slici. Na primjer, okidač može biti bijeli pravokutnik pozicioniran u donjem lijevom kutu slike. Naravno, korišteni uzorak može biti proizvoljne kompleksnosti i veličine. Kod napada BadNets, izmjene oznaka su najčešće tipa *all-to-one*, ali česte su i izmjene tipa *all-to-all*.



**Slika 2.1.** Primjer primjene napada BadNets. Izvornoj slici (lijevo) dodaje se okidač kako bi nastala zatrovana slika (desno).

### 2.1.2. Napad Blend

Napad Blend provodi miješanje zadanog okidača sa svakim odabranim ulaznim primjerom. Pritom je jačina napada određena hiperparametrom  $\alpha$  koji nazivamo jačina miješanja (engl. *blending strength*). Primjenu napada Blend možemo prikazati jednačbom:

$$\tilde{x} = (1 - \alpha) \cdot x + \alpha \cdot t \quad (2.1)$$

Pri čemu  $x$  označava ulazni primjer,  $t$  okidač, a  $\tilde{x}$  zatrovani primjer. Kod napada Blend, izmjene oznaka su uobičajeno tipa *all-to-one*.



**Slika 2.2.** Primjer primjene napada Blend. Izvorna slika (lijevo) miješa se s okidačem uz  $\alpha = 0.2$  kako bi nastala zatrovana slika (desno).

### 2.1.3. Napad WaNet

Napad WaNet provodi geometrijsku transformaciju svakog odabranog ulaznog primjera koristeći nasumično generirano deformacijsko polje. Deformacijsko polje svakom pikselu određene slike dodjeljuje vektor pomaka prema pikselu izvorne slike. Pritom hiperparametar  $k$  određuje veličinu nasumično generiranog polja šuma na temelju kojeg se skaliranjem i interpolacijom dobiva konačno deformacijsko polje, a hiperparametar  $s$  određuje jačinu deformacije. Primjenu napada WaNet možemo definirati jednačbom:

$$\tilde{x} = \mathcal{W}(x, \mathbf{M}(k, s)) \quad (2.2)$$

Pri čemu  $\mathbf{x}$  označava ulazni primjer,  $\mathbf{M}$  deformacijsko polje generirano uz hiperparametre  $k$  i  $s$ ,  $\mathcal{W}$  primjenu deformacijskog polja na ulazni primjer, a  $\tilde{\mathbf{x}}$  zatrovani primjer. Kao i kod napada Blend, kod napada WaNet su izmjene oznaka uobičajeno tipa *all-to-one*.



**Slika 2.3.** Primjer primjene napada WaNet. Izvorna slika (lijevo) transformira se koristeći deformacijsko polje uz  $k = 8$  i  $s = 4$  kako bi nastala zatrovana slika (desno). Hiperparametri  $k$  i  $s$  su uvećani kako bi učinak trovanja bio uočljiviji.

## 2.2. Primjeri algoritama za obranu od zatrovanih podataka

U ovome radu, rezultate okvira VIBE uspoređujemo s rezultatima triju algoritama za obranu od zatrovanih podataka: *Anti-backdoor learning* (ABL) [5], *Decoupling based defense* (DBD) [6] te *Adaptively splitting dataset-based defense* (ASD) [7].

### 2.2.1. Obrana ABL

Algoritam *Anti-backdoor learning* (ABL) sastoji se od dva glavna koraka: izoliranje zatrovanih podataka (engl. *backdoor isolation*) i odučavanje trovanja (engl. *backdoor unlearning*). Osnovna ideja ove obrane je da se nakon određenog broja epoha učenja uz posebno definiran gubitak izolira određeni broj primjera za koje se smatra da su zatrovani. Nakon prvog koraka, ti se primjeri koriste za odučavanje trovanja, dok se preostali primjeri koriste za standardno učenje.

Konkretno, cilj prvog koraka je zadržati vrijednost gubitka svakog pojedinog primjera oko praga  $\gamma$ . Kako bi ovo postigli, autori predlažu korištenje gradijentnog uspona u slučaju da gubitak primjera padne ispod praga, dok se inače koristi gradijentni spust. Gubitak u prvom koraku možemo prikazati jednadžbom:

$$\mathcal{L}_1 = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} [\text{sign}(\ell(f_{\theta}(\mathbf{x}), \mathbf{y}) - \gamma) \cdot \ell(f_{\theta}(\mathbf{x}), \mathbf{y})] \quad (2.3)$$

Pritom  $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$  označava primjer  $\mathbf{x}$  s pripadnom oznakom  $\mathbf{y}$  iz skupa podataka  $\mathcal{D}$ ,  $f_{\theta}(\mathbf{x})$  izlaz modela s parametrima  $\theta$ ,  $\ell(\hat{\mathbf{y}}, \mathbf{y})$  gubitak za predviđenu oznaku  $\hat{\mathbf{y}}$  i stvarnu oznaku  $\mathbf{y}$ , a sign operaciju signum.

Ideja je da će gubitak za zatrovane primjere veoma brzo pasti ispod praga te će se za njih često aktivirati gradijentni uspon, dok će gubitak čistih primjera sporije padati i stabilizirati se oko praga. Nakon zadanog broja epoha, izolira se udio  $p$  primjera s najnižim gubitkom i proglašava potencijalnim zatrovanim skupom. Važno je napomenuti da bismo prvi korak ABL-a mogli zamijeniti proizvoljnim algoritmom detekcije zatrovanih primjera.

U drugom koraku, učenje se u svakoj epohi provodi zasebno za procijenjeni čisti odnosno zatrovani skup. Dok se učenje na čistom skupu provodi uz standardni gradijentni spust, učenje na zatrovanom skupu provodi se uz gradijentni uspon kako bi model odučili od trovanja. Ovo je moguće zato što je trovanje najčešće realizirano uz samo jedan ciljani razred tj. uz *all-to-one* način izmjene oznaka. Gubitak u drugom koraku možemo prikazati jednadžbom:

$$\mathcal{L}_2 = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \hat{\mathcal{D}}_c} [\ell(f_{\theta}(\mathbf{x}), \mathbf{y})] - \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \hat{\mathcal{D}}_b} [\ell(f_{\theta}(\mathbf{x}), \mathbf{y})] \quad (2.4)$$

Pritom  $\hat{\mathcal{D}}_c$  označava procijenjeni čisti skup, a  $\hat{\mathcal{D}}_b$  procijenjeni zatrovani skup.

### 2.2.2. Obrana DBD

Algoritam *Decoupling based defense* (DBD) problem učenja modela na zatrovanim podacima razdvaja na tri koraka. Pritom DBD model tretira kao dvije povezane cjeline: ekstraktor značajki (engl. *feature extractor*) te klasifikator (najčešće nekoliko potpuno povezanih slojeva). Ekstraktor značajki za ulazni primjer na izlazu daje vektor značajki tj. ugrađivanje (engl. *embedding*) u latentnom metričkom prostoru. S druge strane, klasifikator za ugrađivanje na ulazu predviđa jednu od mogućih oznaka.



**Slika 2.4.** Prikaz glavnih koraka obrane DBD. Preuzeto iz [6].

U prvom koraku, DBD uči ekstraktor značajki koristeći proizvoljan algoritam samonadziranog učenja [17] na slikama bez oznaka. Budući da algoritmi samonadziranog učenja uobičajeno uključuju korištenje jakih augmentacija ulaznih primjera, kvaliteta okidača kod zatrovanih primjera bit će narušena. Dodatno, jer se model u ovoj fazi uči bez oznaka, u prvom koraku nije moguće zatrovati ekstraktor značajki. Drugim riječima, na kraju prvog koraka imat ćemo naučen čisti ekstraktor značajki.

U drugom koraku, parametri ekstraktora značajki su zamrznuti, a klasifikator učimo koristeći klasično nadzirano učenje na podacima s oznakama. Pritom kao funkciju gubitka koristimo simetričnu unakrsnu entropiju - pokazano je da korištenje iste rezultira višim gubitkom kod zatrovanih primjera [18]. Ipak, ako bismo koristili samo ova dva koraka, dobiveni model ne bi imao performanse jednake stanju tehnike jer smo ekstraktor značajki učili bez oznaka. Zbog toga, važno je napraviti podjelu skupa podataka na čisti i zatrovani skup te nakon toga ugoditi (engl. *fine-tune*) cijeli model. Kod algoritma DBD, ova podjela se radi na temelju iznosa gubitka: udio  $\alpha$  primjera s najnižim iznosom gubitka smatrat ćemo vjerodostojnim tj. čistim skupom, dok ćemo preostale primjere smatrati zatrovanima.

Treći korak koristi podjelu na čisti i zatrovani skup kako bi dodatno ugodio parametre cijelog modela. Konkretno, zatrovanom skupu uklanjamo oznake te potom model učimo koristeći proizvoljni algoritam polunadziranog učenja [19]. Na kraju ovog koraka, imat ćemo naučen cjelokupni model otporan na zatrovane primjere.

### 2.2.3. Obrana ASD

Algoritam *Adaptively splitting dataset-based defense* (ASD) problem učenja modela na zatrovanim podacima razdvaja na tri koraka. Tijekom sva tri koraka, održavaju se dva skupa podataka: čisti i zagađeni skup. Pritom se u čistom skupu nalaze podatci za koje je velika vjerojatnost da su čisti, dok se u zagađenom skupu nalaze zatrovani podatci, kao i preostali čisti podatci. Kroz učenje, čisti skup se povećava dodavanjem primjera iz zagađenog skupa. Konačno, na kraju učenja bi zagađeni skup trebao sadržavati isključivo zatrovane podatke. Parametri modela uvijek se ažuriraju na temelju proizvoljnog polunadziranog gubitka, pri čemu se zagađeni skup koristi za učenje bez oznaka.



**Slika 2.5.** Prikaz glavnih koraka obrane ASD.  $D_C$  predstavlja čisti skup, a  $D_P$  zagađeni skup.  $L_1$  odgovara gubitku  $\mathcal{L}_{SCE}$ , a  $L_2$  gubitku  $\mathcal{L}_{CE}$ . Preuzeto iz [7].

U prvom koraku, čisti skup se inicijalizira na mali broj provjereno čistih primjera (na primjer, po 10 primjera za svaki razred), dok se zagađeni skup inicijalizira na cijeli skup podataka. Svakih  $t$  epoha učenja, u čisti skup se iz zagađenog skupa dodaje  $n$  primjera iz svakog pojedinog razreda koji imaju najniži gubitak  $\mathcal{L}_{SCE}$ . Pritom kao funkciju gubitka  $\mathcal{L}_{SCE}$  koristimo simetričnu unakrsnu entropiju kako bi zatrovani primjeri imali što viši gubitak. Ovu podjelu zovemo razredno-svjesna podjela vođena gubitkom (engl. *class-aware loss-guided split*).

Tijekom drugog koraka, čisti skup značajno proširujemo dodavanjem udjela  $\alpha$  zagađenog skupa. Pritom se dodaju primjeri iz cijelog skupa (neovisno o razredu) koji imaju najniži gubitak  $\mathcal{L}_{SCE}$ . Ovu podjelu zovemo razredno-nesvjesna podjela vođena gubitkom (engl. *class-agnostic loss-guided split*).

Nakon prva dva koraka ASD-a, u zagađenom skupu preostaju zatrovani primjeri, ali i neki teški primjeri specifični za model. Zbog toga što model nismo učili na teškim primjerima s oznakama, performanse modela trenutno su niže od stanja tehnike. Kako bismo dodali i teške primjere u čisti skup, svaku epohu trećeg koraka konstruiramo virtualni model. Virtualni model inicijaliziramo parametrima glavnog modela te potom provodimo jednu epohu nadziranog učenja na zagađenom skupu uz korištenje gubitka  $\mathcal{L}_{CE}$ . Kao funkciju gubitka  $\mathcal{L}_{CE}$  koristimo standardnu unakrsnu entropiju. Nakon učenja virtualnog modela, mjerimo smanjenje gubitka definirano jednadžbom:

$$\Delta\mathcal{L}_{SCE} = \mathcal{L}_{SCE}(f_{\theta}) - \mathcal{L}_{SCE}(f_{\theta'}) \quad (2.5)$$

Pritom  $f_{\theta}$  označava glavni model s parametrima  $\theta$ , a  $f_{\theta'}$  predstavlja virtualni model s parametrima  $\theta'$  dobivenim nakon jedne epohe nadziranog učenja na zagađenom skupu. Konačno, udio  $\gamma$  primjera s najmanjim smanjenjem gubitka dodajemo u čisti skup. Intuicija iza ove podjele je da su zatrovani podatci lagani za naučiti, tako da već nakon jedne epohe nadziranog učenja isti imaju veoma nizak gubitak, dok teški čisti primjeri i dalje imaju visok gubitak. Važno je napomenuti da se virtualni model koristi isključivo za provođenje podjele na temelju smanjenja gubitka. Ovu podjelu zovemo meta-podjela (engl. *meta-split*) jer je pristup s učenjem virtualnog modela inspiriran područjem meta-učenja (engl. *meta-learning*) [20]. Na kraju trećeg koraka, u čistom skupu će se nalaziti gotovo svi čisti primjeri, a dobiveni model će imati izvrsne performanse uz otpornost na zatrovane primjere.

### 3. Problem zašumljenih oznaka

Proces stvaranja skupa vizualnih podataka otprilike možemo podijeliti u tri koraka. U prvom koraku, potrebno je definirati skup razreda tj. oznaka koje mogu biti dodijeljene svakom primjeru. Tijekom drugog koraka, cilj nam je prikupiti što veći skup slika, pri tom pazeći na to da slike precizno prikazuju situacije s kojima će se naš model susretati. Dodatno, veoma je važno da je prikupljeni skup što raznovrsniji kako bi model mogao naučiti dobro generalizirati. Konačno, u trećem koraku anotatori označavaju slike na temelju definiranog skupa oznaka. Drugi i treći korak ovog procesa mogu se provoditi jedan za drugim, ali i paralelno - nakon što se prikupi određen broj slika, taj skup se šalje na označavanje, a istovremeno je moguće prikupiti još slika.

Iako anotatori podataka prolaze kroz brojne edukacije kako bi se što bolje upoznali s pravilima na temelju kojih će označavati podatke, određeni podatci su prirodom teži od drugih pa može doći do pogrešnog označavanja. Problem zašumljenih oznaka sličan je problemu zatrovanih podataka po tome što se kod oba problema model mora nositi s pogrešnim oznakama. Ipak, kod problema zašumljenih oznaka podrazumijevamo da su ulazne slike netaknute tj. da ne postoji nikakva izmjena ulaza. Dodatno, kod problema zašumljenih oznaka ne postoji konkretan cilj - podatci sa zašumljenim oznakama nastaju slučajno, bez ikakvih loših namjera.

Dok je kod zatrovanih podataka problem prelako učenje poveznice između prisutnosti okidača i određenog ciljnog razreda, kod podataka sa zašumljenim oznakama je problem činjenica da model uz dovoljan kapacitet može naučiti oznake čak i ako su one nasumično generirane [1]. Drugim riječima, ako nismo oprezni, lako je doći do prena-učenosti (engl. *overfitting*) modela. Naravno, ovakav model će loše generalizirati na ispravno označenim podacima.



Kako bismo mogli usporediti performanse različitih algoritama za učenje na zašumljenim oznakama, najčešće određenom udjelu čistog skupa podataka dodajemo sintetičke zašumljene oznake. Hiperparametar koji opisuje udio podataka sa zašumljenim oznakama zvat ćemo stopa šuma (engl. *noise rate*). Nakon učenja na zašumljenim podacima, performanse modela evaluiramo na čistom, ali i na u potpunosti zašumljenom skupu. Cilj algoritama za učenje na zašumljenim oznakama tada je zadržati performanse modela učenog na zašumljenim podacima što bližima performansama modela učenog na čistim podacima.

### 3.1. Primjeri metoda zašumljivanja oznaka

U ovome radu, fokusiramo se na dvije metode zašumljivanja oznaka: simetrično (engl. *symmetric*) i asimetrično (engl. *asymmetric*) zašumljivanje [21].

#### 3.1.1. Simetrično zašumljivanje

Kod simetričnog zašumljivanja, nasumično biramo zadani udio primjera iz čistog skupa podataka te istima dodjeljujemo nasumično generirane nove oznake. Ovu metodu zašumljivanja također zovemo i uniformno zašumljivanje jer svaki razred ima jednaku vjerojatnost postati nova oznaka za neki zadani primjer.

Pritom razlikujemo dvije vrste simetričnog zašumljivanja: inkluzivnu (engl. *symm-inc*) i ekskluzivnu (engl. *symm-exc*) vrstu [21]. Ako govorimo o inkluzivnom simetričnom zašumljivanju, stvarna oznaka za neki zadani primjer može biti odabrana i kao zašumljena oznaka. S druge strane, kod ekskluzivnog simetričnog zašumljivanja ignoriramo stvarnu oznaku tj. zašumljena oznaka se uvijek razlikuje od stvarne oznake primjera. U našem radu, fokusiramo se na inkluzivno simetrično zašumljivanje.



**Slika 3.1.** Prikaz inkluzivnog (gore) odnosno ekskluzivnog (dolje) simetričnog zašumljivanja za tri odabrana primjera iz skupa CIFAR10 [22]. Čiste oznake (lijevo) preslikavaju se na jednu od mogućih oznaka (sredina) i nastaju zašumljene oznake (desno). Vidimo da primjeri s istim čistim oznakama nemaju nužno i iste zašumljene oznake.

### 3.1.2. Asimetrično zašumljivanje

Kod asimetričnog zašumljivanja, prvo nasumično biramo zadani udio primjera zasebno za svaki razred. Drugim riječima, ako je u pitanju skup CIFAR10 [22], za svaki od 10 razreda nasumično ćemo odabrati zadani udio primjera kojima ćemo potencijalno dodijeliti zašumljene oznake.

Nakon početnog odabira, svakom primjeru dodjeljujemo novu oznaku na temelju predodređenog preslikavanja. Pritom je moguće da određeni razredi nemaju definirano preslikavanje u neki drugi razred, tako da odabranim primjerima iz tih razreda neće biti dodijeljene zašumljene oznake. Preslikavanje na temelju kojeg se pojedinim primjerima dodjeljuje nova oznaka unaprijed je definirano za pojedine skupove podataka poput skupova MNIST [23], CIFAR10 i CIFAR100 [22]. Kod skupova s hijerarhijskim odnosom razreda poput skupa CIFAR100, preslikavanja su definirana nasumično unutar pojedinih grupa razreda.



**Slika 3.2.** Prikaz preslikavanja razreda za skup CIFAR10. Odabranim primjerima se na temelju čistih oznaka (lijevo) dodjeljuju zašumljene oznake (desno). Pritom odabrani primjeri s istim čistim oznakama uvijek imaju i iste zašumljene oznake.

## 3.2. Primjeri algoritama za učenje na zašumljenim oznakama

U ovome radu, rezultate prilagođenog okvira VIBE uspoređujemo s rezultatima dvaju algoritama za učenje na podacima sa zašumljenim oznakama: *Sparse over-parameterization* (SOP) [8] te *Imprecise label learning* (ILL) [9].

### 3.2.1. Algoritam SOP

Osnovna ideja algoritma *Sparse over-parameterization* (SOP) bazira se na činjenici da su primjeri sa zašumljenim oznakama u većini slučajeva rijetki. Jedan od mogućih načina za učenje modela na zašumljenim oznakama tada je modeliranje šuma za pojedine primjere koristeći rijetke vektore.

Konkretno, za svaki primjer  $(\mathbf{x}_i, \mathbf{y}_i)$  iz skupa podataka  $\mathcal{D}$  definiramo rijetki vektor  $\mathbf{s}_i$  koji služi kao odstupanje između uočene (potencijalno zatrovane) oznake  $\mathbf{y}_i$  i čiste (skrivenne) oznake  $\mathbf{l}_i$ . Uobičajeno, učenje modela se svodi na potragu za parametrima  $\theta$  koji minimiziraju gubitak definiran jednažbom:

$$\mathcal{L} = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} [\ell(f_{\theta}(\mathbf{x}), \mathbf{y})] \quad (3.1)$$

Pritom  $f_{\theta}$  predstavlja model s parametrima  $\theta$ , a  $\ell(\hat{\mathbf{y}}, \mathbf{y})$  funkciju gubitka između predviđanja modela  $\hat{\mathbf{y}}$  i dane oznake  $\mathbf{y}$ . Uz uvođenje rijetkog vektora  $\mathbf{s}_i$  za svaki primjer, učenje se svodi na potragu za parametrima  $(\theta, \{\mathbf{s}_i\}_{i=1}^N)$  koji minimiziraju gubitak:

$$\mathcal{L}_{SOP} = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} [\ell(f_{\theta}(\mathbf{x}) + \mathbf{s}, \mathbf{y})] \quad (3.2)$$

Dodatno, kako bi se osigurala rijetkost vektora  $\mathbf{s}_i$ , isti definiramo pomoću vektora  $\mathbf{u}_i$  i vektora  $\mathbf{v}_i$  kao:

$$\mathbf{s}_i = \mathbf{u}_i \odot \mathbf{u}_i - \mathbf{v}_i \odot \mathbf{v}_i \quad (3.3)$$

Drugim riječima, za svaki podatak  $(\mathbf{x}_i, \mathbf{y}_i)$  definiramo dva vektora parametara:  $\mathbf{u}_i$  i  $\mathbf{v}_i$ . Ove vektore učimo zajedno s parametrima modela  $\theta$  koristeći gradijentni spust. Pritom je stopa učenja za parametre  $\{\mathbf{u}_i, \mathbf{v}_i\}_{i=1}^N$  skalirana hiperparametrom  $\alpha$  kako bi se izbjeglo trivijalno rješenje optimizacijskog problema kod kojeg vrijedi  $\mathbf{u}_i \equiv \mathbf{v}_i \equiv \mathbf{0}$ ,  $\forall i \in N$ .

Uočimo da vektor  $\mathbf{s}_i$  za primjer  $(\mathbf{x}_i, \mathbf{y}_i)$  dodatno mora poštovati nekoliko uvjeta: u vektoru  $\mathbf{s}_i$  se pozitivna vrijednost može nalaziti isključivo na mjestu koje odgovara nenegativnoj vrijednosti u  $\mathbf{y}_i$ , dok se negativne vrijednosti mogu nalaziti isključivo na mjestima koje odgovaraju vrijednosti 0 u  $\mathbf{y}_i$ . Dodatno, svaki element vektora  $\mathbf{s}_i$  mora biti u rasponu  $[-1, 1]$ . Kako bismo ovo osigurali, uvodimo konačnu definiciju vektora  $\mathbf{s}_i$ :

$$\mathbf{s}_i = \mathbf{u}_i \odot \mathbf{u}_i \odot \mathbf{y}_i - \mathbf{v}_i \odot \mathbf{v}_i \odot (\mathbf{1} - \mathbf{y}_i) \quad (3.4)$$

Pritom su vektori parametara  $\mathbf{u}_i$  i  $\mathbf{v}_i$  definirani kao:

$$\mathbf{u}_i \in [-1, 1]^K, \mathbf{v}_i \in [-1, 1]^K \quad (3.5)$$

U algoritmu SOP, za učenje parametara  $(\theta, \{\mathbf{u}_i\}_{i=1}^N)$  se kao funkcija gubitka koristi standardna unakrsna entropija. Nažalost, unakrsna entropija se ne može koristiti za ispravno učenje parametara  $\{\mathbf{v}_i\}_{i=1}^N$ . Umjesto toga, za učenje tih parametara kao funkciju gubitka koristimo srednju kvadratnu pogrešku.

Kako bi poboljšali performanse osnovnog algoritma SOP, autori predlažu dodavanje dvije nove komponente gubitka: konzistencijski gubitak (engl. *consistency loss*) [24] te gubitak ravnoteže razreda (engl. *class-balance loss*) [25]. Inačicu algoritma koja dodatno koristi ove dvije komponente gubitka zovemo SOP+, a ista općenito postiže bolje performanse od osnovne inačice algoritma.

### 3.2.2. Algoritam ILL

Algoritam *Imprecise label learning* (ILL) obuhvaća i pokušava riješiti nekoliko problema vezanih uz nesavršenosti oznaka: učenje na djelomičnim oznakama (engl. *partial label learning*) [26], polunadzirano učenje te učenje na zašumljenim oznakama. Konkretno, ILL sve ove probleme smatra vrstama problema nepreciznih oznaka (engl. *imprecise labels*). Skup podataka tada sadrži ulazne slike  $X$  i neprecizne oznake  $I$ , dok su čiste oznake  $Y$  skrivene (latentne). Pritom su neprecizne oznake apstraktne - u stvarnosti to može biti skup oznaka (u problemu učenja na djelomičnim oznakama), ali i zašumljena oznaka (u problemu učenja na zašumljenim oznakama). Cilj učenja modela tada je pronaći parametre  $\theta$  koji maksimiziraju zajedničku vjerojatnost definiranu kao:

$$p(X, I | \theta) = \sum_Y p(X, I, Y | \theta) \quad (3.6)$$

Kako bismo maksimizirali logaritam očekivanja uz prisutnost latentne varijable, koristimo algoritam maksimizacije očekivanja (algoritam EM) [27]. Pritom u E koraku algoritma maksimizacije očekivanja računamo očekivanje zajedničke vjerojatnosti  $p(X, I, Y | \theta)$  uz zadanu uvjetnu vjerojatnost  $p(Y | X, I, \theta^t)$  u vremenskom koraku  $t$ . S druge strane, u M koraku tražimo parametre  $\theta$  koji maksimiziraju donju varijacijsku granicu (engl. *evidence lower bound* - ELBO) vjerojatnosti  $p(X, I | \theta)$ . Budući da se različite vrste problema razlikuju primarno po prirodi nepreciznosti oznaka, algoritmi za pojedine probleme najviše će se razlikovati u načinu izračuna vjerojatnosti  $p(Y | X, I, \theta^t)$ . Ovime algoritam ILL pruža unificirani okvir koji podržava različite vrste nepreciznosti oznaka, kao i kombinacije istih.

Dakle, glavni cilj algoritma ILL je pronaći parametre  $\theta$  koji maksimiziraju zajedničku vjerojatnost  $p(X, I | \theta)$  odnosno logaritam iste. Prema algoritmu EM, ovaj problem možemo riješiti maksimizacijom očekivanja:

$$\mathbb{E}_{Y|X,I,\theta^t} [\log p(X, Y, I | \theta)] = \mathbb{E}_{Y|X,I,\theta^t} [\log p(Y | X, \theta) + \log p(I | X, Y, \theta)] \quad (3.7)$$

Pritom ignoriramo član  $p(X)$  jer isti ne ovisi o parametrima  $\theta$ . Iz dobivenog izraza vidimo da izračunom očekivanja razmatramo sve moguće oznake  $Y$  na temelju danih nepreciznih oznaka  $I$ , umjesto da se fokusiramo na samo jednu prepravljenu oznaku. Na temelju ove formulacije dalje možemo izvesti cilj za pojedini problem. U našem radu, fokusirat ćemo se na izvod za problem učenja na zašumljenim oznakama.

Kod problema učenja na zašumljenim oznakama, neprecizne oznake  $I$  se manifestiraju kao zašumljene oznake  $\hat{Y}$ . Vjerojatnost  $p(\hat{Y}|Y, X, \theta)$  tada predstavlja model šuma ovisan o pojedinom uzorku  $X$ . Ovaj problem pojednostavit ćemo razmatranjem modela šuma neovisnog o uzorku  $\mathcal{T}(\hat{Y}|Y, \omega)$  s parametrima  $\omega$ . Cilj koji želimo maksimizirati tada dobivamo kao:

$$\begin{aligned}
\mathbb{E}_{Y|X,I,\theta^t} [\log p(X, Y, I|\theta)] &= \mathbb{E}_{Y|X,I,\theta^t} [\log p(Y, I|X, \theta)] \\
&= \mathbb{E}_{Y|X,\hat{Y},\theta^t} [\log p(Y, \hat{Y}|X, \theta)] \\
&= \mathbb{E}_{Y|X,\hat{Y},\theta^t} [\log p(Y|\hat{Y}, X, \theta) + \log p(\hat{Y}|X, \theta)] \\
&= \sum_Y p(Y|\hat{Y}, X, \theta^t) \log p(Y|\hat{Y}, X, \theta) + \log p(\hat{Y}|X, \theta)
\end{aligned} \tag{3.8}$$

Funkcija gubitka koju želimo minimizirati tada postaje:

$$\mathcal{L}_{ILL} = \mathcal{L}_{CE}(p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{y}}, \theta, \omega^t), p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{y}}, \theta^t, \omega^t)) + \mathcal{L}_{CE}(p(\hat{\mathbf{y}}|\mathbf{x}, \theta, \omega), \hat{\mathbf{y}}) \tag{3.9}$$

Pritom  $\mathcal{L}_{CE}$  označava unakrsnu entropiju. Prva komponenta dobivenog gubitka odgovara konzistencijskom gubitku čistih oznaka uvjetovanih zašumljenim oznakama, dok druga komponenta odgovara nadziranom gubitku na zašumljenim oznakama. Pritom se za izračun obje komponente koristi model šuma uz zadanu zašumljenu oznaku  $\hat{\mathbf{y}}$ :

$$\begin{aligned}
p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{y}}, \theta, \omega^t) &\propto p(\mathbf{y}|\mathbf{x}, \theta) \mathcal{T}(\hat{\mathbf{y}}|\mathbf{y}, \omega^t) \\
p(\hat{\mathbf{y}}|\mathbf{x}, \theta, \omega) &= \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \theta) \mathcal{T}(\hat{\mathbf{y}}|\mathbf{y}, \omega)
\end{aligned} \tag{3.10}$$

## 4. Samonadzirano učenje

Samonadzirano učenje [17] paradigma je strojnog učenja kod koje model uči izlučivati korisne reprezentacije tj. značajke ulaznih podataka na temelju posebno osmišljenih zadataka bez oznaka (engl. *pretext tasks*). Ovim pristupom pokušava se adresirati problem cijene i vremena potrebnog za označavanje velikih skupova podataka. Model učen proizvoljnim algoritmom samonadziranog učenja dalje se može koristiti kao okosnica (engl. *backbone*) za model koji rješava neki nizvodni (engl. *downstream*) zadatak poput klasifikacije ili detekcije objekata. Pritom se okosnici dodaje klasifikacijska glava (najčešće nekoliko potpuno-povezanih slojeva), a cijeli model se dodatno ugađa nadziranim učenjem na manjem skupu podataka prilagođenom konkretnom problemu.

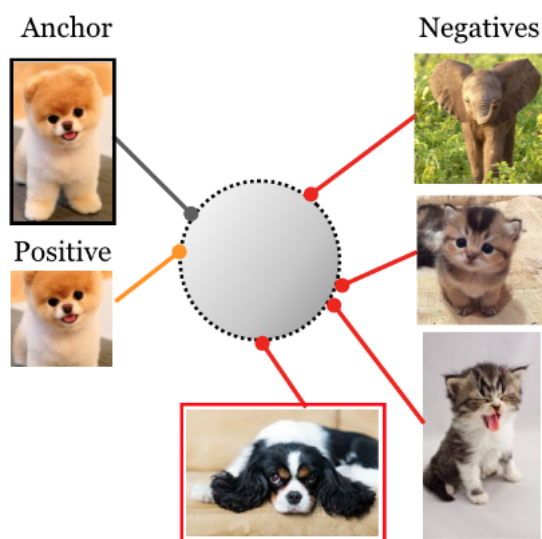
Ključno pitanje kod samonadziranog učenja je formiranje zadatka učenja bez oznaka tj. odlučivanje o tome na temelju čega će model dobivati signal za učenje. Tipični zadatci uključuju rekonstrukciju ulaznih podataka na temelju izlučenih reprezentacija, grupiranje reprezentacija tj. ugrađivanja (engl. *embedding*) semantički sličnih podataka u prostoru ugrađivanja ili predviđanje maskiranih piksela ulaznih slika (engl. *masked image modeling* - MIM) [28]. Rješavanjem jednog od zadataka učenja, model posredno uči izlučivati korisne reprezentacije ulaznih podataka ili uočavati korisne odnose između podataka.

Područje samonadziranog učenja možemo podijeliti na temelju korištenog tipa zadatka učenja, a neka od najpoznatijih područja su autoasocijativno samonadzirano učenje [29] i kontrastno samonadzirano učenje [17]. Kod autoasocijativnog samonadziranog učenja, osnovni zadatak je rekonstruirati ulazni podatak na temelju izlučene reprezentacije. Pritom je model koji učimo odgovoran za izlučivanje reprezentacije, dok za rekonstrukciju uobičajeno koristimo drugi model koji nakon učenja odbacujemo. U našem radu, fokusirat ćemo se na kontrastno samonadzirano učenje.

## 4.1. Kontrastno samonadzirano učenje

Kontrastno učenje jedno je od područja samonadziranog učenja. Ono podrazumijeva učenje izlučivanja korisnih reprezentacija tj. ugrađivanja ulaznih podataka na temelju parova podataka. Ako su dobivena ugrađivanja normirana, a sličnost dvaju ugrađivanja možemo izračunati koristeći neku od standardnih metrika (na primjer, kosinusnu sličnost), tada govorimo o metričkim ugrađivanjima tj. ugrađivanjima u metrički prostor [30]. Možemo reći da naučeni model preslikava ulazne slike na  $(d-1)$  dimenzionalnu hipersferu  $S^{d-1}$ , pri čemu sličnost ugrađivanja odgovara semantičkoj sličnosti ulaza.

Kod kontrastnog učenja razlikujemo sidro, pozitivne i negativne primjere. Trenutno promatrani podatak iz minigrupe nazivamo sidro, podatak sličan sidru nazivamo pozitivan primjer, a podatak različit od sidra nazivamo negativan primjer. Budući da primjeri nisu označeni, pozitivne primjere najčešće dobivamo perturbacijom sidra, dok negativnim primjerima smatramo sve ostale podatke iz minigrupe.



**Slika 4.1.** Prikaz osnovne ideje kontrastnog samonadziranog učenja. Cilj učenja je približiti sidro (gore lijevo) i pozitivan primjer (dolje lijevo), ali i međusobno udaljiti sidro i negativne primjere (desno). Preuzeto iz [31].

Glavni cilj kontrastnog učenja je približiti ugrađivanja pozitivnih parova (sidra i pozitivnog primjera), ali i istovremeno udaljiti ugrađivanja negativnih parova (sidra i negativnih primjera). Kako bismo ovo postigli, veoma je važno prikladno definirati funkciju gubitka, a neke od mogućih su trojni gubitak [32] te gubitak N parova. Gubitak N parova također je poznat i kao infoNCE gubitak [33], a možemo ga definirati jednadžbom:



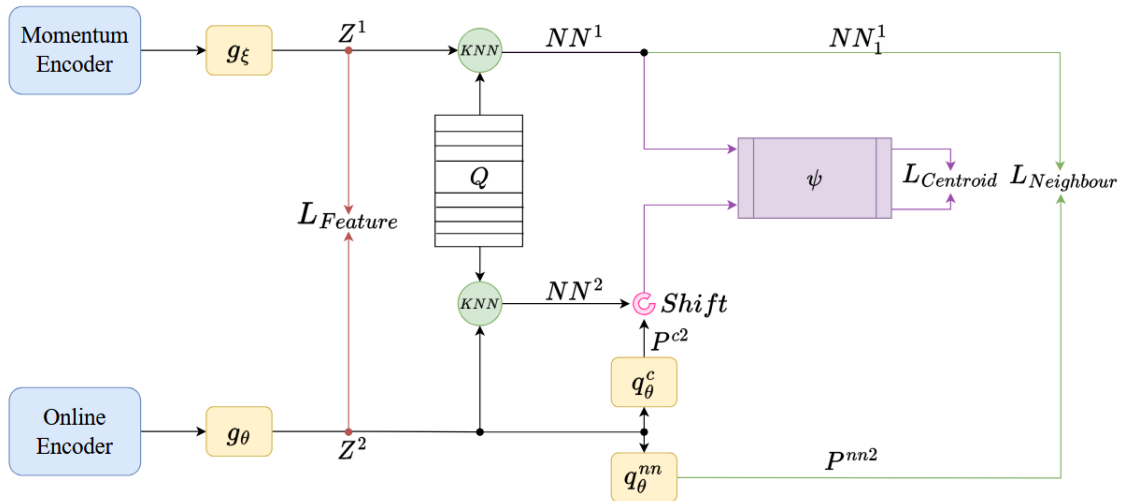
$$\mathcal{L}_{infoNCE} = -\log \frac{\exp(\langle \mathbf{z}_a, \mathbf{z}_p \rangle / \tau)}{\sum_{i=1}^N \exp(\langle \mathbf{z}_a, \mathbf{z}_{ni} \rangle / \tau)} \quad (4.1)$$

Pritom  $\mathbf{z}_a$  označava ugrađivanje sidra  $\mathbf{x}_a$ ,  $\mathbf{z}_p$  ugrađivanje pozitivnog primjera  $\mathbf{x}_p$ ,  $\mathbf{z}_{ni}$  ugrađivanje i-tog negativnog primjera  $\mathbf{x}_{ni}$  iz minigrupe, a  $\tau$  hiperparametar temperature. Oznaka  $\langle \dots \rangle$  označava skalarni produkt vektora unutar zagrada.

U algoritmu za obranu od zatrovanih podataka VIBE, samonadzirana inicijalizacija parametara modela veoma je važan korak koji osigurava uspješnost obrane. Pritom za inicijalizaciju koristimo okvir All4One [34] koji se bazira na kontrastnom samonadziranom učenju.

## 4.2. Okvir All4One

Okvir All4One kombinira tri zasebna pristupa kontrastiranju kako bi naučeni modeli postizali što bolje rezultate. Konkretno, optimizacijski cilj okvira All4One sadrži komponentu kontrastiranja najbližih susjeda (engl. *nearest neighbour contrast*) [35], komponentu kontrastiranja centroida (engl. *centroid contrast*) te komponentu kontrastiranja značajki (engl. *feature contrast*) [36].



**Slika 4.2.** Prikaz arhitekture okvira All4One. Komponenta kontrastiranja najbližih susjeda označena je zelenom bojom, komponenta kontrastiranja centroida ljubičastom bojom, a komponenta kontrastiranja značajki crvenom bojom. Preuzeto iz [34].

U okviru All4One, model koji učimo služi kao koder značajki  $f_\theta$ . Osim glavnog koda koji učimo gradijentnim spustom, tijekom učenja se dodatno održava i koder s momentom  $f_\xi$  čiji parametri odgovaraju eksponencijalnom pomičnom prosjeku (engl. *exponential moving average* - EMA) parametara glavnog koda. Dodatno, osim para koda, učimo i par projektora  $g_\theta$  i  $g_\xi$ . Za svaku sliku  $\mathbf{x}$  na ulazu, generiraju se dvije augmentirane slike  $\mathbf{x}^1$  te  $\mathbf{x}^2$ . Jedna od dobivenih augmentiranih slika tada prolazi kroz koder s momentom i pripadni projektor te dobivamo ugrađivanje  $\mathbf{z}^1 = g_\xi(f_\xi(\mathbf{x}^1))$ , dok druga augmentirana slika prolazi kroz glavni koder i pripadni projektor te dobivamo ugrađivanje  $\mathbf{z}^2 = g_\theta(f_\theta(\mathbf{x}^2))$ . Dobivena ugrađivanja dalje se mogu koristiti za izračun pojedinih komponenti optimizacijskog cilja. Osim para koda i para projektor, dodatno se održavaju i dva prediktora  $q_\theta^{nn}$  te  $q_\xi^c$  koji služe za prilagodbu ugrađivanja za zadatak kontrastiranja najbližih susjeda odnosno zadatak kontrastiranja centroida.

#### 4.2.1. Kontrastiranje najbližih susjeda

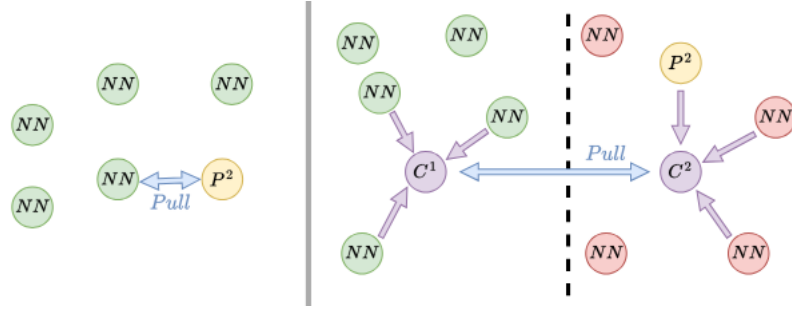
Kontrastiranje najbližih susjeda često je korišten pristup kontrastnog samonadziranog učenja kod kojega se ugrađivanje jedne od perturbiranih slika zamijeni sa svojim najbližim susjedom. Konkretno, tijekom učenja održavamo skup prethodnih ugrađivanja (engl. *support set*)  $Q$  koji možemo koristiti za pronalazak najbližih susjeda za neko zadano ugrađivanje. Ovu operaciju označit ćemo operatorom  $KNN$ . Gubitak za  $i$ -ti primjer  $\mathbf{x}_i$  u minigrupi tada možemo definirati jednadžbom:

$$\mathcal{L}_{NNCLR} = -\log \frac{\exp(\langle \mathbf{nn}_i^1, \mathbf{p}_i^{nn,2} \rangle / \tau)}{\sum_{k=1}^N \exp(\langle \mathbf{nn}_i^1, \mathbf{p}_k^{nn,2} \rangle / \tau)} \quad (4.2)$$

Pritom  $\mathbf{nn}_i^1$  označava najbližeg susjeda ugrađivanja prve perturbacije primjera  $\mathbf{x}_i$ ,  $\mathbf{p}_i^{nn,2}$  izlaz prediktora  $q_\theta^{nn}$  za ugrađivanje druge perturbacije primjera  $\mathbf{x}_i$ , a  $\tau$  hiperparametar temperature. Naspram osnovne inačice kontrastnog učenja, kontrastiranje najbližih susjeda povećava raznolikost primjera koje model vidi tijekom učenja, a time i generalizacijsku moć naučenog modela.

### 4.2.2. Kontrastiranje centroida

Ipak, pokazano je da modeli ućeni kontrastiranjem najbliŹih susjeda mogu postići još bolje performanse ako se pritom koristi više od jednog najbliŹeg susjeda [37]. NaŹalost, ućenje s više od jednog najbliŹeg susjeda znaćajno povećava trajanje ućenja jer se funkcija cilja treba evaluirati za svakog susjeda pojedinaćno. Autori okvira All4One ovome problemu doskaću uvođenjem kontrastiranja centroida.



**Slika 4.3.** Prikaz kontrastiranja najbliŹih susjeda (lijevo) i kontrastiranja centroida (desno). Kod kontrastiranja najbliŹih susjeda, kontrastiraju se najbliŹi susjed jedne perturbacije i druga perturbacija. S druge strane, kod kontrastiranja centroida, kontrastiraju se centroidi pojedinih perturbacija. Preuzeto iz [34].

Konkretno, nakon Źto smo za zadani primjer  $x$  generirali ugrađivanja  $z^1$  te  $z^2$ , za ista prvo pronalazimo skupove najbliŹih susjeda  $NN^1$  odnosno  $NN^2$ . Ugrađivanje  $z^2$  tada prolazi kroz prediktor  $q_\theta^c$  te dobivamo ugrađivanje  $p^{c,2}$ . Ovim ugrađivanjem zamijenit ćemo posljednjeg najbliŹeg susjeda u skupu  $NN^2$  te potom provesti operaciju kruŹnog posmaka (engl. *shift*) kojim će ugrađivanje  $p^{c,2}$  doći na prvo mjesto u skupu.

Kada imamo pripremljene skupove najbliŹih susjeda  $NN^1$  te  $NN^2$ , iz istih Źelimo dobiti po jedno ugrađivanje koje sadrŹi informacije o svim najbliŹim susjedima iz pripadnog skupa. Kako bismo ovo postigli, koristimo mehanizam samopaŹnje (engl. *self-attention*) [38]. Konkretno, koristimo koder transformera  $\psi$  kojemu na ulaz dovodimo niz ugrađivanja  $Seq$  iz skupa  $NN$  obogaćen sinusoidalnim pozicijskim kodiranjem [38]. Na izlazu koderu  $\psi$  tada dobivamo niz ugrađivanja  $Seq^c$  obogaćenih kontekstualnim informacijama o preostalim susjedima iz skupa. Kako bismo za svaki skup imali samo jedno ugrađivanje, kao centroid  $c$  proglašavamo prvo ugrađivanje s izlaza koderu transformera  $Seq_1^c$ .

Ovaj postupak provodimo zasebno za skup  $NN^1$  odnosno skup  $NN^2$  kako bismo dobili centroide  $\mathbf{c}^1$  odnosno  $\mathbf{c}^2$ . Konačno, gubitak za  $i$ -ti primjer  $\mathbf{x}_i$  u minigrupi možemo definirati jednadžbom:

$$\mathcal{L}_{centroid} = -\log \frac{\exp(\langle \mathbf{c}_i^1, \mathbf{c}_i^2 \rangle / \tau)}{\sum_{k=1}^N \exp(\langle \mathbf{c}_i^1, \mathbf{c}_k^2 \rangle / \tau)} \quad (4.3)$$

### 4.2.3. Kontrastiranje značajki

Kontrastiranje značajki veoma se razlikuje od uobičajenih pristupa kontrastnom učenju. Kod ovog pristupa, ideja je direktno kontrastirati značajke ugrađivanja. Kako bismo ovo postigli, potrebno je izračunati korelacijsku matricu značajki za svaku minigrupu. Tada je cilj naučiti parametre modela koji korelacijsku matricu značajki što više približavaju prema jediničnoj matrici.

Konkretno, prvo za svaku minigrupu konstruiramo matrice  $\mathbf{Z}^1$  odnosno  $\mathbf{Z}^2$  u kojima reci označavaju ugrađivanja prve odnosno druge perturbacije pripadnih primjera iz minigrupe. Nakon provođenja  $L_2$  normalizacije matrice po stupcima, korelacijsku matricu  $\mathbf{C}\mathbf{C}^1$  dobivamo računanjem kosinusne sličnosti između transponirane matrice  $\mathbf{Z}^1$  te matrice  $\mathbf{Z}^2$ . Dodatno, korelacijsku matricu  $\mathbf{C}\mathbf{C}^2$  dobivamo uz zamjenu grana tj. zamjenu glavnog kodera i kodera s momentom te ponovan izračun sličnosti. Kada smo izračunali obje korelacijske matrice, gubitak za zadanu minigrupu možemo definirati jednadžbom:

$$\mathcal{L}_{feature} = \frac{1}{2} \sqrt{\frac{1}{2D} \sum_{i=1}^D ((1 - \mathbf{C}\mathbf{C}_{ii}^1)^2 + (1 - \mathbf{C}\mathbf{C}_{ii}^2)^2)} + \frac{1}{2} \sqrt{\frac{1}{2D(D-1)} \sum_{i=1}^D \sum_{j \neq i}^D ((\mathbf{C}\mathbf{C}_{ij}^1)^2 + (\mathbf{C}\mathbf{C}_{ij}^2)^2)} \quad (4.4)$$

Pritom  $D$  označava broj značajki tj. dimenzionalnost vektora ugrađivanja. Vidimo da će gubitak biti to manji što su korelacijske matrice bliže jediničnoj matrici. Prvi član gubitka pokušava povećati invarijantnost pojedinih značajki na augmentacije ulaznih slika, dok drugi član pokušava smanjiti međusobnu redundantnost značajki.

#### 4.2.4. Kombinirano kontrastiranje

Dok prethodni algoritmi samonadziranog učenja većinom koriste samo jednu vrstu kontrastiranja (na primjer, samo kontrastiranje najbližih susjeda), okvir All4One kombinira sva tri navedena pristupa. Konkretno, gubitak koji želimo minimizirati možemo definirati jednadžbom:

$$\mathcal{L}_{All4One} = \sigma \mathcal{L}_{NNCLR} + \kappa \mathcal{L}_{centroid} + \eta \mathcal{L}_{feature} \quad (4.5)$$

Pritom su  $\sigma$ ,  $\kappa$  i  $\eta$  hiperparametri koji određuju jačinu utjecaja pojedine vrste kontrastiranja. Kombiniranjem više pristupa kontrastnog učenja, okvir All4One poboljšava učenje korisnih reprezentacija ulaznih podataka.

## 5. Algoritam maksimizacije očekivanja

Zamislamo da imamo skup opaženih varijabli  $X$  te želimo pronaći parametre  $\theta$  pretpostavljenog modela za koje je log-izglednost  $\log p(X|\theta)$  maksimalna. Drugim riječima, zanima nas MLE (engl. *maximum likelihood estimation*) procjena [39] parametara  $\theta$ . Osim što ovaj zadatak ponekad ima rješenje u zatvorenoj formi (engl. *closed-form solution*), općenito ga možemo riješiti gradijentnim postupcima poput stohastičkog gradijentnog spusta [40]. Ako uz opažene varijable  $X$  postoje i skrivene (latentne) varijable  $Z$ , problem pronalaska MLE procjene parametara  $\theta$  postaje kompliciraniji. Konkretno, nepotpunu log-izglednost parametara tada možemo definirati kao:

$$\log p(X|\theta) = \log \int p(X, Z|\theta) dZ = \log \int p(X|Z, \theta) p(Z|\theta) dZ \quad (5.1)$$

Gledajući da optimizacijski cilj sada sadrži varijable  $Z$  čije vrijednosti ne znamo, više ga nije moguće direktno maksimizirati. Jedan od mogućih algoritama koji se nosi s postojanjem skrivenih varijabli je algoritam maksimizacije očekivanja (engl. *expectation maximization algorithm* - algoritam EM) [27]. Kao i gradijentni spust, i algoritam EM je iterativne prirode, ali se pritom svaka iteracija sastoji od dva zasebna koraka: E koraka i M koraka.

Pokazuje se da algoritam EM osigurava monotoni rast nepotpune log-izglednosti parametara  $\log p(X|\theta)$ , ali pritom ne postoji garancija da će algoritam konvergirati u globalni maksimum [41]. Drugim riječima, moguće je da dobiveni parametri  $\theta$  odgovaraju lokalnom maksimumu nepotpune log-izglednosti, tako da se preporučuje koristiti heurističke pristupe poput ponovnog pokretanja algoritma s nasumično odabranim početnim vrijednostima parametara  $\theta$ .

## 5.1. Osnovna formulacija algoritma

Osnovna ideja algoritma EM je jednostavna. Tijekom učenja, u svakoj iteraciji prvo provodimo E korak, a nakon njega provodimo M korak. Gledajući da ne znamo vrijednosti skrivenih varijabli  $Z$ , kao ni vrijednosti optimalnih parametara  $\theta$ , učenje možemo započeti uz nasumično odabrane vrijednosti parametara  $\theta^{(0)}$ .

Na početku E koraka, fiksiramo trenutne vrijednosti parametara  $\theta^{(t)}$ . Tada računamo očekivanje potpune log-izglednosti  $\log p(X, Z|\theta)$  uz uvjetnu distribuciju skrivenih varijabli  $p(Z|X, \theta^{(t)})$ . Izračunato očekivanje možemo definirati jednadžbom:

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z \sim p(\cdot|X, \theta^{(t)})} [\log p(X, Z|\theta)] \quad (5.2)$$

U M koraku, cilj nam je pronaći parametre  $\theta$  koji maksimiziraju izračunato očekivanje  $Q(\theta|\theta^{(t)})$ . Nakon što smo izračunali nove parametre  $\theta^{(t+1)}$ , možemo ih iskoristiti za provođenje E koraka sljedeće iteracije. Općenito, jednu iteraciju algoritma EM sažeto možemo prikazati kao:

$$\theta^{(t+1)} = \arg \max_{\theta} \mathbb{E}_{Z \sim p(\cdot|X, \theta^{(t)})} [\log p(X, Z|\theta)] \quad (5.3)$$

Pokažimo sada općenitu formulaciju algoritma maksimizacije očekivanja.

## 5.2. Općenita formulacija algoritma

## **6. Transportni problem**



## **7. VIBE**

## **8. Skup podataka**

## **9. Eksperimenti**

## **10. Zaključak**

## Literatura

- [1] C. Zhang, S. Bengio, M. Hardt, B. Recht, i O. Vinyals, “Understanding deep learning requires rethinking generalization”, *arXiv preprint arXiv:1611.03530*, 2016.
- [2] A. Voulodimos, N. Doulamis, A. Doulamis, i E. Protopapadakis, “Deep learning for computer vision: A brief review”, *Computational intelligence and neuroscience*, sv. 2018, br. 1, str. 7068349, 2018.
- [3] B. Biggio, B. Nelson, i P. Laskov, “Poisoning attacks against support vector machines”, *arXiv preprint arXiv:1206.6389*, 2012.
- [4] S. Gupta i A. Gupta, “Dealing with noise problem in machine learning data-sets: A systematic review”, *Procedia Computer Science*, sv. 161, str. 466–474, 2019.
- [5] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, i X. Ma, “Anti-backdoor learning: Training clean models on poisoned data”, *Advances in Neural Information Processing Systems*, sv. 34, str. 14 900–14 912, 2021.
- [6] K. Huang, Y. Li, B. Wu, Z. Qin, i K. Ren, “Backdoor defense via decoupling the training process”, *arXiv preprint arXiv:2202.03423*, 2022.
- [7] K. Gao, Y. Bai, J. Gu, Y. Yang, i S.-T. Xia, “Backdoor defense via adaptively splitting poisoned dataset”, u *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023., str. 4005–4014.
- [8] S. Liu, Z. Zhu, Q. Qu, i C. You, “Robust training under label noise by over-parameterization”, u *International Conference on Machine Learning*. PMLR, 2022., str. 14 153–14 172.

- [9] H. Chen, A. Shah, J. Wang, R. Tao, Y. Wang, X. Li, X. Xie, M. Sugiyama, R. Singh, i B. Raj, “Imprecise label learning: A unified framework for learning with various imprecise label configurations”, *Advances in Neural Information Processing Systems*, sv. 37, str. 59 621–59 654, 2024.
- [10] I. Sabolić, M. Grcić, i S. Šegvić, “Seal your backdoor with variational defense”, *arXiv preprint arXiv:2503.08829*, 2025.
- [11] T. Gu, K. Liu, B. Dolan-Gavitt, i S. Garg, “Badnets: Evaluating backdooring attacks on deep neural networks”, *IEEE Access*, sv. 7, str. 47 230–47 244, 2019.
- [12] Y. Chen, X. Gong, Q. Wang, X. Di, i H. Huang, “Backdoor attacks and defenses for deep neural networks in outsourced cloud environments”, *IEEE Network*, sv. 34, br. 5, str. 141–147, 2020.
- [13] A. Nguyen i A. Tran, “Wanet-imperceptible warping-based backdoor attack”, *arXiv preprint arXiv:2102.10369*, 2021.
- [14] Y. Li, Y. Li, B. Wu, L. Li, R. He, i S. Lyu, “Invisible backdoor attack with sample-specific triggers”, u *Proceedings of the IEEE/CVF international conference on computer vision*, 2021., str. 16 463–16 472.
- [15] K. D. Doan, Y. Lao, i P. Li, “Marksman backdoor: Backdoor attacks with arbitrary target class”, *Advances in Neural Information Processing Systems*, sv. 35, str. 38 260–38 273, 2022.
- [16] M. Barni, K. Kallas, i B. Tondi, “A new backdoor attack in cnns by training set corruption without label poisoning”, u *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019., str. 101–105.
- [17] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, i F. Makedon, “A survey on contrastive self-supervised learning”, *Technologies*, sv. 9, br. 1, str. 2, 2020.
- [18] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, i J. Bailey, “Symmetric cross entropy for robust learning with noisy labels”, u *Proceedings of the IEEE/CVF international conference on computer vision*, 2019., str. 322–330.

- [19] J. E. Van Engelen i H. H. Hoos, “A survey on semi-supervised learning”, *Machine learning*, sv. 109, br. 2, str. 373–440, 2020.
- [20] R. Vilalta i Y. Drissi, “A perspective view and survey of meta-learning”, *Artificial intelligence review*, sv. 18, str. 77–95, 2002.
- [21] F. R. Cordeiro i G. Carneiro, “A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations?” u *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE, 2020., str. 9–16.
- [22] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images”, 2009.
- [23] L. Deng, “The mnist database of handwritten digit images for machine learning research [best of the web]”, *IEEE signal processing magazine*, sv. 29, br. 6, str. 141–142, 2012.
- [24] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, i C. A. Raffel, “Mixmatch: A holistic approach to semi-supervised learning”, *Advances in neural information processing systems*, sv. 32, 2019.
- [25] D. Tanaka, D. Ikami, T. Yamasaki, i K. Aizawa, “Joint optimization framework for learning with noisy labels”, u *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018., str. 5552–5560.
- [26] Y. Tian, X. Yu, i S. Fu, “Partial label learning: Taxonomy, analysis and outlook”, *Neural Networks*, sv. 161, str. 708–734, 2023.
- [27] T. K. Moon, “The expectation-maximization algorithm”, *IEEE Signal processing magazine*, sv. 13, br. 6, str. 47–60, 1996.
- [28] V. Hondru, F. A. Croitoru, S. Minaee, R. T. Ionescu, i N. Sebe, “Masked image modeling: A survey”, *arXiv preprint arXiv:2408.06687*, 2024.
- [29] M. A. Kramer, “Nonlinear principal component analysis using autoassociative neural networks”, *AIChE journal*, sv. 37, br. 2, str. 233–243, 1991.

- [30] E. Chávez, G. Navarro, R. Baeza-Yates, i J. L. Marroquín, “Searching in metric spaces”, *ACM computing surveys (CSUR)*, sv. 33, br. 3, str. 273–321, 2001.
- [31] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, i D. Krishnan, “Supervised contrastive learning”, *Advances in neural information processing systems*, sv. 33, str. 18 661–18 673, 2020.
- [32] F. Schroff, D. Kalenichenko, i J. Philbin, “Facenet: A unified embedding for face recognition and clustering”, u *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015., str. 815–823.
- [33] A. v. d. Oord, Y. Li, i O. Vinyals, “Representation learning with contrastive predictive coding”, *arXiv preprint arXiv:1807.03748*, 2018.
- [34] I. G. Estepa, I. Sarasúa, B. Nagarajan, i P. Radeva, “All4one: Symbiotic neighbour contrastive learning via self-attention and redundancy reduction”, u *Proceedings of the IEEE/CVF international conference on computer vision*, 2023., str. 16 243–16 253.
- [35] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, i A. Zisserman, “With a little help from my friends: Nearest-neighbor contrastive learning of visual representations”, u *Proceedings of the IEEE/CVF international conference on computer vision*, 2021., str. 9588–9597.
- [36] J. Zbontar, L. Jing, I. Misra, Y. LeCun, i S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction”, u *International conference on machine learning*. PMLR, 2021., str. 12 310–12 320.
- [37] S. A. Koohpayegani, A. Tejankar, i H. Pirsiavash, “Mean shift for self-supervised learning”, u *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021., str. 10 326–10 335.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, i I. Polosukhin, “Attention is all you need”, *Advances in neural information processing systems*, sv. 30, 2017.
- [39] I. J. Myung, “Tutorial on maximum likelihood estimation”, *Journal of mathematical Psychology*, sv. 47, br. 1, str. 90–100, 2003.



- [40] S.-i. Amari, “Backpropagation and stochastic gradient descent method”, *Neurocomputing*, sv. 5, br. 4-5, str. 185–196, 1993.
- [41] C. J. Wu, “On the convergence properties of the em algorithm”, *The Annals of statistics*, str. 95–103, 1983.

## Sažetak

### Varijacijsko učenje na zašumljenim oznakama

Dominik Jambrović

Sažetak...

**Ključne riječi:** prva ključna riječ; druga ključna riječ; treća ključna riječ